
**GenPPI: Um Software Autônomo Para
Predição *Ab Initio* de Redes de Interação Entre
Proteínas Bacterianas**

William Ferreira dos Anjos



UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Uberlândia
2021

William Ferreira dos Anjos

**GenPPI: Um Software Autônomo Para
Predição *Ab Initio* de Redes de Interação Entre
Proteínas Bacterianas**

Dissertação de mestrado apresentada ao Programa de Pós-graduação da Faculdade de Computação da Universidade Federal de Uberlândia como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Ciência da Computação

Orientador: Prof. Dr. Anderson Rodrigues dos Santos

Uberlândia

2021

Ficha Catalográfica Online do Sistema de Bibliotecas da UFU
com dados informados pelo(a) próprio(a) autor(a).

A599 2021	<p>Anjos, William Ferreira dos, 1986- GenPPI: um software autônomo para predição ab initio de redes de interação entre proteínas bacterianas [recurso eletrônico] / William Ferreira dos Anjos. - 2021.</p> <p>Orientador: Anderson Rodrigues dos Santos. Dissertação (Mestrado) - Universidade Federal de Uberlândia, Pós-graduação em Ciência da Computação. Modo de acesso: Internet. Disponível em: http://doi.org/10.14393/ufu.di.2021.365 Inclui bibliografia. Inclui ilustrações.</p> <p>1. Computação. I. Santos, Anderson Rodrigues dos, 1971- , (Orient.). II. Universidade Federal de Uberlândia. Pós-graduação em Ciência da Computação. III. Título.</p> <p style="text-align: right;">CDU: 681.3</p>
--------------	--

Bibliotecários responsáveis pela estrutura de acordo com o AACR2:

Gizele Cristine Nunes do Couto - CRB6/2091



ATA DE DEFESA - PÓS-GRADUAÇÃO

Programa de Pós-Graduação em:	Ciência da Computação				
Defesa de:	Mestrado Acadêmico, 13/2021, PPGCO				
Data:	30 de junho de 2021	Hora de início:	14h	Hora de encerramento:	16h30min
Matrícula do Discente:	11822CCP012				
Nome do Discente	William Ferreira dos Anjos				
Título do Trabalho:	GENPPI: um software autônomo para predição ab initio de redes de interação entre proteínas bacterianas				
Área de concentração:	Ciência da Computação				
Linha de pesquisa:	Inteligência Artificial				
Projeto de Pesquisa de vinculação:	-				

Reuniu-se, por videoconferência, a Banca Examinadora, designada pelo Colegiado do Programa de Pós-graduação em Ciência da Computação, assim composta: Professores Doutores: Laurence Rodrigues Amaral - FACOM/UFU; Marcos Augusto dos Santos - UFMG e Anderson Rodrigues dos Santos - FACOM/UFU, orientador do candidato.

Os examinadores participaram desde as seguintes localidades: Marcos Augusto dos Santos - Belo Horizonte/MG; Laurence Rodrigues Amaral - Patos de Minas/MG e Anderson Rodrigues dos Santos - Uberlândia/MG. O discente participou da cidade de Uberlândia/MG.

Iniciando os trabalhos o presidente da mesa, Prof. Dr. Anderson Rodrigues dos Santos, apresentou a Comissão Examinadora e o candidato, agradeceu a presença do público, e concedeu ao Discente a palavra para a exposição do seu trabalho. A duração da apresentação do Discente e o tempo de arguição e resposta foram conforme as normas do Programa.

A seguir o senhor presidente concedeu a palavra, pela ordem sucessivamente, aos examinadores, que passaram a arguir o candidato. Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando o candidato:

Aprovado.

Esta defesa faz parte dos requisitos necessários à obtenção do título de Mestre.

O competente diploma será expedido após cumprimento dos demais requisitos, conforme as normas do Programa, a legislação pertinente e a regulamentação interna da UFU.

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente ata que após lida e achada conforme foi assinada pela Banca Examinadora.





Documento assinado eletronicamente por **Anderson Rodrigues dos Santos, Professor(a) do Magistério Superior**, em 01/07/2021, às 11:57, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **Marcos Augusto dos Santos, Usuário Externo**, em 10/08/2021, às 12:21, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2870295** e o código CRC **FC9BC51A**.

*Dedico este trabalho a todos aqueles a quem esta
pesquisa possa ajudar de alguma forma.*

Agradecimentos

Agradeço principalmente a Deus pela vida e por até aqui ter me ajudado em todos os meus caminhos.

Agradeço à minha família, em especial minha mãe Neusa Maria dos Anjos por tamanha dedicação, fruto de seu grande amor pelos filhos.

Não poderia deixar de agradecer ao meu orientador Professor Anderson Rodrigues dos Santos por ter acreditado em mim me concedendo a oportunidade de realizar este trabalho e de obter essa conquista.

Aos professores da Pós Graduação da FACOM/UFU, àqueles com quem tive a oportunidade de estudar, agradeço pelo aprendizado.

Agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela contribuição financeira.

Também quero agradecer aos secretários da pós-graduação, Erisvaldo e Sônia, por terem sido sempre gentis e prestativos auxiliando-me em processos burocráticos inerentes ao curso.

Finalizo agradecendo a todos amigos e colegas que contribuíram diretamente ou indiretamente agregando valores e conhecimentos na minha formação.

“A gravidade explica os movimentos dos planetas, mas não pode explicar quem colocou os planetas em movimento. Deus governa todas as coisas e sabe tudo que é ou que pode ser feito.”

(Isaac Newton)

Resumo

As interações proteína-proteína (do inglês, Protein-Protein Interactions – PPI) desempenham um papel fundamental na determinação do resultado da maioria dos processos celulares. Identificar corretamente as interações de proteínas e as redes de PPI que elas compreendem, é de fundamental importância para o entendimento dos mecanismos moleculares dentro da célula. Isso pode fornecer informações úteis na realização de tarefas críticas como a fabricação de drogas e vacinas contra doenças causadas por agentes infecciosos. Abordagens computacionais são utilizadas combinando várias fontes de dados biológicos, a fim de prever interações de proteínas com níveis satisfatórios de confiabilidade. Neste trabalho, propõe-se um novo software autônomo de bioinformática (GenPPI) para predição *ab initio* de redes de interação entre proteínas bacterianas. A solução proposta analisa genomas buscando por evidências de eventos evolutivos que indicam interações de proteínas. A saber, eventos de vizinhança gênica conservada, fusão gênica e perfil filogenético conservado. Este trabalho também introduz uma nova heurística para comparação par-a-par de sequências de aminoácidos de proteínas. Como resultados, primeiramente demonstra-se a eficácia da heurística proposta comparando sua exatidão com o BLASTp, o principal algoritmo heurístico para comparação de sequências proteicas. A exatidão do dois algoritmos heurísticos é estimada verificando qual se aproxima mais do algoritmo exato Needleman-Wunsh, utilizado para comparação de sequências biológicas. A heurística proposta superou o BLASTp apresentando maior exatidão na comparação par-a-par de proteínas e menor tempo de processamento. Posteriormente, a confiabilidade biológica das predições computacionais realizadas, é verificada. Para tanto, foram feitas análises de filogenia a partir de dados gerados pelo programa, após processar genomas de gêneros bacterianos selecionados como estudos de caso. Foram analisados 28 genomas do gênero *Dietzia*, 45 de *Rhodococcus*, 50 de *Corynebacterium* e 81 de *Aeromonas*. As análises de filogenia realizadas demonstram correção e confiabilidade biológica para as redes de interação proteica preditas pelo software desenvolvido. Por final, compara-se a qualidade de redes de interação geradas pelo GenPPI com uma rede do STRING, a principal ferramenta do estado da arte deste trabalho. Tal compara-

ção mostra que a solução proposta é capaz de gerar redes de tão boa qualidade quanto as redes do STRING. Vale mencionar que, com essa solução, é suprida uma deficiência identificada no estado da arte, a indisponibilidade de ferramentas computacionais para prever PPIs sem negligenciar proteínas inéditas. O software desenvolvido encontra-se disponível para download no site: <<https://genppi.facom.ufu.br/>> ou no repositório: <<https://github.com/santosardr/genppi>>, onde também contém um guia do usuário.

Palavras-chave: Biologia Computacional, Bioinformática, Predição *ab initio* de Redes de PPI, Redes Complexas, Heurísticas Para Comparação de Sequências de Proteínas.

Abstract

Protein-protein interactions play a key role in determining the outcome of most cellular processes. Correctly identifying protein interactions and the PPI networks they comprise is of fundamental importance for understanding the molecular mechanisms within the cell. This can provide useful insights in performing critical tasks such as manufacturing drugs and vaccines against diseases caused by infectious agents. Computational approaches are used combining various sources of biological data in order to predict protein interactions with satisfactory levels of reliability. In this work, we propose a new autonomous bioinformatics software (GenPPI) for *ab initio* prediction of interaction networks between bacterial proteins. The proposed solution analyzes genomes looking for evidence of evolutionary events that indicate protein interactions. Namely, conserved gene neighborhood events, gene fusion and conserved phylogenetic profile. This work also introduces a new heuristic for pairwise comparison of protein amino acid sequences. As a result, we first demonstrate the effectiveness of the proposed heuristic by comparing its accuracy with BLASTp, the main heuristic algorithm for comparing protein sequences. The accuracy of the two heuristic algorithms is estimated by checking which one is closest to the exact Needleman-Wunsh algorithm used for comparing biological sequences. The proposed heuristic surpassed BLASTp, presenting greater accuracy in the pair-by-pair comparison of proteins and shorter processing time. Subsequently, the biological reliability of the computational predictions performed is verified. Therefore, phylogeny analyzes were performed using data generated by the program, after processing genomes of bacterial genera selected as case studies. 28 genomes of the genus *Dietzia*, 45 of *Rhodococcus*, 50 of *Corynebacterium* and 81 of *Aeromonas* were analyzed. The phylogeny analyzes performed demonstrate correctness and biological reliability for the protein interaction networks predicted by the developed software. Finally, the quality of interaction networks generated by GenPPI is compared with a STRING network, the main state-of-the-art tool of this work. This comparison shows that the proposed solution is capable of generating networks of as good quality as the STRING networks. It is worth mentioning that, with this solution, a deficiency identified in the state of the art, the unavailability of computational tools to

predict PPIs without neglecting new proteins, is addressed. The developed software is available for download on the site: <<https://genppi.facom.ufu.br/>> or on the repository: <<https://github.com/santosardr/genppi>>, where also contains a user guide.

Keywords: Computational Biology, Bioinformatics, *Ab Initio* Prediction of PPI Networks, Complex Networks, Heuristics for Protein Sequence Comparison.

Lista de ilustrações

Figura 1 – Rede de interação proteína-proteína do fungo <i>Saccharomyces Cerevisae</i>	29
Figura 2 – Dogma central da biologia molecular	38
Figura 3 – As quatro estruturas de uma proteína	40
Figura 4 – Ilustração do pan-genoma de três cepas fictícias	42
Figura 5 – Vértices adjacentes	43
Figura 6 – Grau dos vértices. O vértice 1 tem duas arestas incidentes, portanto grau 2. Os vértices 2 e 3 tem uma aresta incidente, portanto, ambos tem grau 1. Assim, o grau máximo desse grafo é 2 e o grau mínimo é 1.	43
Figura 7 – A diferença entre dois grafos com base em sua direcionalidade. A figura à esquerda mostra um grafo não direcionado, enquanto a figura à direita mostra um grafo direcionado.	44
Figura 8 – Exemplo de um grafo ponderado	44
Figura 9 – Exemplo de um grafo completo ou totalmente conexo	45
Figura 10 – Ilustração da predição de PPIs baseada em eventos de vizinhança gênica conservada. Infere-se que as proteínas A, B e C interagem entre si, pois elas ocorreram repetidamente próximas umas das outras (no mesmo <i>operon</i>) em pelo menos 2 do conjunto de 4 genomas analisados.	53
Figura 11 – Ilustração da predição de PPI baseada em eventos de fusão gênica. Infere-se que as proteínas A e B interagem entre si, pois foram fundidas em uma única proteína (proteína Rosetta Stone) em outro genoma.	53
Figura 12 – Ilustração da predição de PPI baseada em eventos de perfil filogenético conservado. Infere-se que as proteínas A e C interagem entre si, pois possuem o mesmo perfil filogenético (10101). Isto é, elas ocorrem nos mesmos genomas e estão ausentes nos mesmos genomas de referência.	55
Figura 13 – Fluxograma do GenPPI	64
Figura 14 – Uma proteína em um arquivo multi- <i>fasta</i>	65
Figura 15 – Ilustração da tabela hash <i>genomas</i>	67

Figura 16 – Exemplo de dois registros da tabela-hash <i>pangenoma</i> , referentes a um par de proteínas classificadas como sendo similares pela heurística	71
Figura 17 – Exemplo de um registro da tabela-hash <i>pangenoma</i>	72
Figura 18 – Ilustração da expansão fixa com uma janela de tamanho igual a 10	74
Figura 19 – Ilustração da primeira configuração de conservação gênica	75
Figura 20 – Ilustração da segunda configuração de conservação gênica	75
Figura 21 – Exemplificação de redundância na criação de arestas de PPI ao considerar tanto a vizinhança gênica à esquerda quanto à direita de uma proteína conservada	77
Figura 22 – Ilustração da expansão dinâmica	79
Figura 23 – Ilustração da janela de expansão dinâmica final	80
Figura 24 – Interação proteica por fusão gênica	81
Figura 25 – Perfil filogenético conservado	85
Figura 26 – Ilustração de perfis semelhantes	86
Figura 27 – Trecho de um arquivo de interação proteína-proteína gerado após uma execução do GenPPI para analisar 5 linhagens da bactéria <i>Buchnera Aphidicola</i>	86
Figura 28 – Gráfico Boxplot Sobre o Perfil Filogenético das Proteínas Conservadas de Cada Genoma (Genome)	105
Figura 29 – Gráfico de calor sobre a similaridade das espécies de bactérias analisadas, com base em seu pan-genoma	107
Figura 30 – Gráfico boxplot da quantidade de genes conservados na vizinhança gênica dos genes recorrentes, considerando uma janela de expansão fixa de tamanho igual a 7	108
Figura 31 – Gráfico Boxplot Sobre o Perfil Filogenético das Proteínas Conservadas de Cada Genoma (Genome)	110
Figura 32 – Gráfico de calor sobre a similaridade das espécies de bactérias analisadas, com base em seu pan-genoma	112
Figura 33 – Gráfico boxplot da quantidade de genes conservados na vizinhança gênica dos genes recorrentes, considerando uma janela de expansão fixa de tamanho igual a 7	114
Figura 34 – Gráfico boxplot da quantidade de genes conservados na vizinhança gênica dos genes recorrentes, considerando uma janela dinâmica com intervalo igual a 3.	115
Figura 35 – Gráfico Boxplot Sobre o Perfil Filogenético das Proteínas Conservadas de Cada Genoma (Genome)	118
Figura 36 – Gráfico de calor sobre a similaridade das espécies de bactérias analisadas, com base em seu pan-genoma	119

Figura 37 – Gráfico boxplot da quantidade de genes conservados na vizinhança gênica dos genes recorrentes, considerando uma janela de expansão fixa de tamanho igual a 7	120
Figura 38 – Gráfico de calor sobre o percentual de proteínas compartilhadas. Consultamos quantos por cento das interações de uma rede listada numa coluna, encontram-se em redes referenciadas nas linhas. O resultado compara as interações em comum entre diferentes redes do GenPPI e a rede do STRING. A intersecção mais significativa do GenPPI obteve 46% das interações do STRING (IDs f1 e d1). Por outro lado, o resultado mais significativo da rede do STRING em relação a nossas redes, atingiu 14%.	123
Figura 39 – Uma rede de interações proteicas gerada pelo GenPPI para o genoma <i>Corynebacterium pseudotuberculosis strain:Ft_2193/67</i> . Os locais subcelulares das proteínas (vértices), representados pelo esquema de cores adotado, são o citoplasma (branco), a membrana (verde), potencialmente expostas na superfície (laranja) e secretadas (azul). Vértices maiores se destacam de acordo com a métrica Bridging Centrality. Esta rede foi desenhada pelo software GEPHI executando os algoritmos de distribuição de dados, nesta ordem: Yifan Hu Multilevel, Fruchterman-Reingold e layouts Force Atlas.	127
Figura 40 – Criação de um diretório para execução do programa	145
Figura 41 – Parâmetros para utilização do GenPPI	146
Figura 42 – Exemplo de uma execução do GenPPI	147
Figura 43 – Diretórios criados em decorrência de uma execução do programa	147
Figura 44 – Trecho do relatório sobre o número de PPIs preditas pelas métricas de vizinhança gênica conservada, perfil filogenético e fusão gênica, para genomas de um conjunto sob análise	154
Figura 45 – Trecho do relatório sobre o pan-genoma de espécies incluídas em uma análise	154
Figura 46 – Trecho do relatório sobre a conservação da vizinhança gênica de genes recorrentes, em uma janela de expansão fixa de tamanho 10	155
Figura 47 – Trecho do relatório sobre interações preditas pelo método de fusão gênica.155	155
Figura 48 – Trecho do relatório que fornece informações sobre o perfil filogenético das proteínas conservadas de genomas analisados. Esse relatório é gerado apenas quando se utiliza o parâmetro -ppcomplete	156
Figura 49 – Gráfico de barras sobre a quantidade de interações proteicas inferidas pelo método de vizinhança gênica conservada.	160
Figura 50 – Histograma da quantidade de perfis filogenéticos encontrados nos genomas	161

Figura 51 – Gráfico de barras sobre o total de interações proteicas inferidas pelo método de perfil filogenético conservado	162
Figura 52 – Gráfico de barras sobre a quantidade de interações proteicas inferidas pelo método de vizinhança gênica conservada.	164
Figura 53 – Histograma da quantidade de perfis filogenéticos encontrados nos genomas	165
Figura 54 – Gráfico de barras sobre o total de interações proteicas inferidas pelo método de perfil filogenético conservado	166
Figura 55 – Gráfico de barras sobre a quantidade de interações proteicas inferidas pelo método de vizinhança gênica conservada.	168
Figura 56 – Histograma da quantidade de perfis filogenéticos encontrados nos genomas	169
Figura 57 – Gráfico de barras sobre o total de interações proteicas inferidas pelo método de perfil filogenético conservado	170

Lista de tabelas

Tabela 1	– Tabela dos 20 aminoácidos.	39
Tabela 2	– Número de genes e interações conhecidas disponíveis para consulta em bancos de dados. Os dados fornecidos nesta tabela foram colhidos nos sites desses banco de dados em Outubro de 2020.	50
Tabela 3	– Heurística para comparação par-a-par de proteínas.	67
Tabela 4	– Fusão gênica via soma dos histogramas de aminoácidos de duas proteínas (A e B).	82
Tabela 5	– Variações de recursos de otimização de código Lisp, aplicados a uma função que gera a sequência de Fibonacci	89
Tabela 6	– Quantidades de ciclos executados pelo processador nas 5 variações de implementações da função de Fibonacci	90
Tabela 7	– Resultados das versões do programa sem otimização de código e com otimização.	98
Tabela 8	– Resultados obtidos com a combinação de diferentes valores para os parâmetros <i>d-limite</i> e <i>qtd-amin</i> da heurística do GenPPI.	100
Tabela 9	– Resultados do GenPPI e BLASTp frente aos resultados do algoritmo exato Needleman-Wunsh.	102
Tabela 10	– Conjunto de parâmetros do GenPPI utilizados na geração de redes de interação para comparação com o STRING.	122
Tabela 11	– Valores de métricas da rede de interação do STRING em comparação com redes obtidas pelo GenPPI.	124
Tabela 12	– Possíveis configurações de parâmetros para a heurística do GenPPI e seus percentuais de identidade mínima garantidos.	149

Lista de siglas

CMNR Corynebacterium, Mycobacterium, Nocardia e Rhodococcus

CN Conserved Neighborhood

GF Gene Fusion

NCBI National Center for Biotechnology Information

PPI Protein-Protein Interactions

PP Phylogenetic Profile

RECOM Rede de Ciências Ômicas

RI Rede de Interação

Sumário

1	INTRODUÇÃO	27
1.1	Motivação	29
1.1.1	O Potencial de Estudos Sobre Redes de Interação Proteína-Proteína	30
1.1.2	Estado da Arte	31
1.2	Objetivos e Desafios da Pesquisa	32
1.2.1	Desafios	32
1.2.2	Objetivo Geral	32
1.2.3	Objetivos Específicos	33
1.3	Hipótese	33
1.4	Contribuições	34
1.5	Organização da Dissertação	34
2	FUNDAMENTAÇÃO TEÓRICA	37
2.1	Fundamentação Biológica	37
2.1.1	Genomas	37
2.1.2	Proteínas	37
2.1.3	Do DNA às Proteínas	38
2.1.4	Estrutura das Proteínas	40
2.1.5	Função das Proteínas	40
2.1.6	Interação Proteína-Proteína	41
2.1.7	Pan-genoma	42
2.2	Fundamentação Matemática e Computacional	42
2.2.1	Grafos	42
2.2.2	Common Lisp	45
2.2.3	Heurística	47
2.2.4	Revisão Literária dos Principais Algoritmos Para Comparação de Sequências Biológicas	48

2.2.5	Revisão Literária de Ferramentas e Métodos Computacionais Para Pre- dição de PPI	49
2.3	Trabalhos Correlatos	57
3	PROPOSTA	61
3.1	Introdução	61
3.2	Tecnologias Utilizadas no Desenvolvimento da Proposta	63
3.3	Fluxograma do GenPPI	63
3.4	Dados de Entrada do GenPPI	65
3.5	Etapa 1 - Gerar Histogramas de Aminoácidos Para Aplicação da Heurística do GenPPI	65
3.6	Etapa 2 - Gerar Pan-genoma	68
3.7	Etapa 3 - Busca de Evidências Biológicas de Interações Pro- teicas no Contexto Genômico	71
3.7.1	Implementação da Predição de PPI Pelo Método de Vizinhança Gênica Conservada	71
3.7.2	Implementação da Predição de PPI Pelo Método de Fusão Gênica	80
3.7.3	Implementação da Predição de PPI Pelo Método de Perfil Filogenético Conservado	83
3.8	Encerramento	86
3.9	Complexidade Algorítmica do GenPPI	87
3.10	Otimização de Código	89
4	EXPERIMENTOS E ANÁLISE DOS RESULTADOS	93
4.1	Métodos de Avaliação	93
4.1.1	Método Para Avaliação de Resultados da Otimização de Código Realizada	93
4.1.2	Método Para Avaliação da Heurística Proposta	94
4.1.3	Método de Validação da Hipótese de Correção Biológica Para Predições de PPI Feitas Pelo Programa Desenvolvido	94
4.1.4	Método Para Comparação de Redes de Interação Geradas Pelo GenPPI Com Uma Rede da Ferramenta STRING	97
4.2	Experimentos	98
4.2.1	Experimentos e Avaliação de Resultados Sobre a Otimização de Código Realizada	98
4.2.2	Experimentos e Avaliação de Resultados Sobre a Heurística Proposta Para Comparação de Sequências Proteicas	99
4.2.3	Experimentos e Avaliação de Resultados Sobre a Hipótese de Correção Biológica Para Predições de PPI Feitas Pelo Programa Desenvolvido	103
4.2.4	Comparação de Redes de Interação Geradas Pelo GenPPI Com Uma Rede do STRING	121

4.2.5	Exemplo de uma Rede de Interação Proteína-Proteína Gerada Pelo GenPPI	126
5	CONCLUSÃO	129
5.1	Principais Contribuições	131
5.2	Trabalhos Futuros	132
5.3	Contribuições em Produção Bibliográfica	133
	REFERÊNCIAS	135

APÊNDICES 143

APÊNDICE A	– GUIA DO USUÁRIO	145
A.1	Parâmetros do GenPPI	148
A.1.1	Parâmetros obrigatórios	148
A.1.2	Parâmetros opcionais	148
A.2	Relatórios Gerados Pelo GenPPI	153
A.2.1	Relatório Sobre o Número de PPIs Preditas Pelos Métodos de Vizinhança Gênica Conservada, Perfil Filogenético, e Fusão Gênica	154
A.2.2	Relatório Sobre o Pan-Genoma	154
A.2.3	Relatório Sobre Vizinhança Gênica Conservada	155
A.2.4	Relatório Sobre Interações Preditas Pelo Método de Fusão Gênica	155
A.2.5	Relatório Sobre o Perfil Filogenético das Proteínas Conservadas de Cada Genoma Incluso em Uma Análise	156
APÊNDICE B	– GRÁFICOS	157
B.1	Gráficos	157
B.1.1	Gráfico de Barras Sobre a Quantidade de Interações Proteicas Inferidas Pela Método de Vizinhança Gênica Conservada	157
B.1.2	Histograma da Quantidade de Perfis Filogenéticos Encontrados nos Genomas	158
B.1.3	Gráfico de Barras Sobre o Total de Interações Proteicas Inferidas pelo Método de Perfil Filogenético Conservado	158
B.2	Resultados Obtidos nos Estudos de Caso Realizados	158
B.2.1	Estudo de Caso 1 – Gênero Bacteriano <i>Dietzia</i>	159
B.2.2	Estudo de Caso 2 – Gênero Bacteriano <i>Corynebacterium</i>	163
B.2.3	Estudo de Caso 3 – Gênero Bacteriano <i>Aeromonas</i>	167

Introdução

Na última década a quantidade de genomas de organismos bacterianos sequenciados e montados aumentou de forma significativa. No mês de Maio do ano 2021 existiam mais de 23 mil genomas completos de procariotos disponíveis para acesso público no site do National Center for Biotechnology Information (NCBI). Nesse site (<<https://www.ncbi.nlm.nih.gov/>>), há também um número muito maior de genomas não finalizados (sem a fita de DNA completa), aproximadamente 306 mil genomas de organismos procariotos. Essa grande quantidade de dados biológicos disponíveis para acesso público atualmente, é consequência de avanços técnicos e tecnológicos que estão ligados às abordagens de sequenciamento de genomas e à bioinformática. Esses dados podem ser utilizados para realizar análises computacionais visando a extração de informações biológicas relevantes.

Apesar da disponibilidade de uma grande quantidade de sequências gênicas de organismos patogênicos aos seres humanos, ainda não conseguimos avanços significativos para combater doenças que possuem esses organismos sequenciados como principais agentes infecciosos. Percebe-se que a elucidação de uma sequência gênica é apenas o primeiro passo na luta contra doenças causadas por agentes microbiológicos patogênicos.

De posse dos milhares de genes, ainda é necessário fazer a seleção de genes promissores para continuar com testes laboratoriais na pesquisa por vacinas. Nesse sentido, considera-se proteínas bacterianas com potencial de utilização na confecção de drogas, vacinas e diagnósticos. Nas células bacterianas existem interações proteicas responsáveis por colonizar e infectar o hospedeiro. Essas interações de proteínas atuam como ponte de comunicação celular entre o vírus e o ser humano, desempenhando um papel vital no processo infeccioso (YANG et al., 2019) (DYER; MURALI; SOBRAL, 2008). Por esta razão, também é necessário investigar as interações entre as proteínas de um patógeno estudado.

Pode-se esperar que uma proteína trabalhe em relativo isolamento, mas espera-se que a maioria opere interagindo umas com as outras para desempenhar funções biológicas específicas nas células (RAMANATHAN; PORTER; KHAVARI, 2019). Isso gera interações par-a-par de proteínas dando origem ao termo: interação proteína-proteína (do

inglês, Protein-Protein Interactions (PPI)).

Identificar *corretamente* os pares de proteínas interagentes de um patógeno, pode auxiliar na descoberta dos mecanismos biológicos pelos quais o mesmo infecta células humanas. Isso pode fornecer insights úteis na realização de tarefas críticas, como a fabricação de drogas e vacinas (YANG et al., 2019). No entanto, os métodos *in vitro* e *in vivo* de identificação de novas interações entre proteínas, são muito caros e demorados. Além disso, apresentam uma taxa de falso-positivos que pode ser bastante elevada gerando erros sistemáticos na detecção das interações de proteínas (BRAUN, 2012). Assim, uma alternativa viável são os métodos *in silico* (computacionais) capazes de prever PPIs biologicamente corretas e confiáveis.

Uma forma de prever computacionalmente as interações entre as proteínas de um organismo bacteriano, é através dos métodos baseados em evidências biológicas no contexto genômico. Esses métodos usam dados genômicos buscando por evidências de eventos evolutivos que indicam interações de proteínas. As proteínas interagentes sofrem certas pressões seletivas durante a evolução das espécies (HIROSE, 2012). Por exemplo, uma pressão para se manterem próximas umas das outras na fita de DNA, a fim de facilitar a geração de seus produtos proteicos. Isso altera a estrutura dos genomas deixando rastros que podem ser detectados ao analisar vários genomas em conjunto (HIROSE, 2012). Existem algumas evidências de eventos evolutivos, que são biologicamente confiáveis para inferir computacionalmente interações de proteínas. Dentre essas, destaca-se:

- **Vizinhança Gênica Conservada:** intervalos recorrentes de genes nos genomas;
- **Fusão Gênica:** quando dois genes individuais de um organismo se fundem como uma sequência única formando um novo gene em outro organismo;
- **Perfil Filogenético Conservado:** genes que ocorrem e estão ausentes nos mesmos genomas de um conjunto de referência.

Tais evidências são exemplos comumente utilizadas na predição computacional de PPIs (ANANTHASUBRAMANIAN et al., 2012) (LIU et al., 2012). As interações proteína-proteína de um organismo podem ser representadas por uma rede complexa, onde os vértices e arestas representam, respectivamente, as proteínas e suas interações. Uma ilustração de rede de interação entre proteínas pode ser vista na Figura 1.

Entretanto, não há muitas alternativas computacionais disponíveis para realizar a predição de redes de PPI de forma *ab initio*¹. A predição *ab initio* é feita a partir dos dados biológicos contidos nos genomas. Um grupo de genomas é analisado visando a busca por evidências biológicas de interação proteica, baseadas nos eventos evolutivos das espécies.

¹ Em biologia computacional, a predição de redes de PPI dita *ab initio* (termo latino que significa “desde o início”), se refere a um processo algorítmico para prever interações de proteínas a partir de dados genômicos.

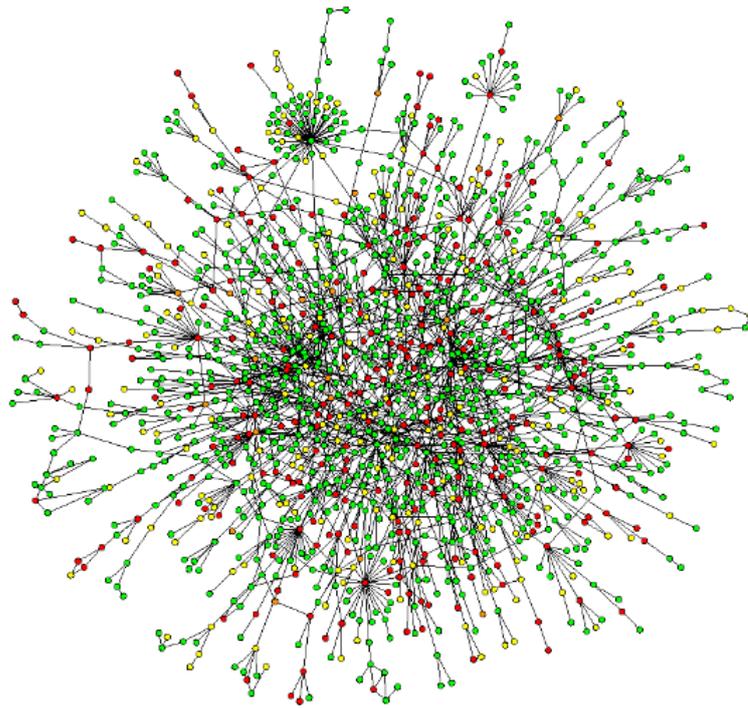


Figura 1 – Rede de interação proteína-proteína do fungo *Saccharomyces Cerevisiae*

Fonte: (JEONG et al., 2001)

Isso inclui evidências de vizinhança gênica conservada, fusão gênica e perfil filogenético conservado, que podem ser detectadas ao analisar vários genomas em conjunto. Uma vantagem da predição *ab initio* em relação às alternativas computacionais predominantes atualmente, é não negligenciar proteínas inéditas na predição de PPI. Essas proteínas costumam representar pelo menos 10% do número total de genes em cada novo genoma sequenciado e montado, não existindo em nenhum outro genoma conhecido (LAPIERRE; GOGARTEN, 2009). Considerando a escassez atual de alternativas computacionais para predição *ab initio* de interações proteicas, e também o constante crescimento do número de genomas completos disponíveis em bancos de dados públicos, este trabalho propõe um novo software autônomo de bioinformática para predição *ab initio* de redes de interação entre proteínas bacterianas.

1.1 Motivação

Nesta seção serão discutidos os fatores motivacionais para o desenvolvimento deste trabalho, sob dois aspectos principais: o potencial de estudos sobre redes de interação de proteínas, e problemas identificados no estado da arte da predição computacional dessas redes.

1.1.1 O Potencial de Estudos Sobre Redes de Interação Proteína-Proteína

As redes de interação de proteínas são de grande valia na compreensão dos processos celulares. Tais redes podem ser úteis na filtragem e avaliação de dados de genômica funcional, além de prover uma plataforma intuitiva para se anotar propriedades funcionais, estruturais e evolutivas de proteínas. Através da exploração das redes de interações de proteínas é possível sugerir novas direções para futuras investigações experimentais.

Atualmente, há muitas aplicações para o estudo das redes de PPI. Muito mais que facilitar a compreensão sobre o complexo sistema biológico de um organismo, essas redes de interação fornecem pistas para identificação de proteínas bacterianas com potencial de utilização na confecção de drogas e vacinas (REZENDE, 2012). Um exemplo recente nesse sentido, é o seu uso na identificação de proteínas relacionadas à infecção por COVID-19. (ZHANG et al., 2020) propuseram um modelo de caminho aleatório para identificar os mecanismos patológicos potenciais de COVID-19 em uma rede de PPI vírus-proteína humana. Efetivamente, os autores identificaram um grupo de proteínas que já foram determinadas como potencialmente importantes para a infecção por COVID-19. Essas proteínas ajudaram no desenvolvimento de drogas e métodos terapêuticos direcionados contra a COVID-19. Em (PUJANA et al., 2007), novos genes associados com maior risco de câncer de mama foram identificados através de análises em redes de interações de proteínas (REZENDE, 2012). Além disso, em parasitologia (especialidade da biologia que estuda os parasitas, os seus hospedeiros e relações entre eles), numa rede de interação entre proteínas do *Plasmodium falciparum*, um dos agentes patológicos da malária, foi identificado um grupo de interações de proteínas relacionado à infecção da doença no hospedeiro (REZENDE, 2012). Também na fitoparasitologia (estudo dos parasitas das plantas), as redes de interação são aplicadas, como por exemplo, o estudo das interações entre as proteínas do fungo *Magnaporthe grise* que provoca uma perda de 10 a 30 por cento da produção anual de arroz (HE et al., 2008). Nesse estudo, os autores conseguiram identificar proteínas envolvidas na patogenicidade do fungo utilizando uma rede de PPI (REZENDE, 2012).

Outra possibilidade com as redes de interação entre proteínas, é a identificação de quais seriam as consequências em se ligar/desligar um determinado gene em um organismo. De posse de uma rede de interação proteica, também é possível realizar predições da função de proteínas para as quais sua função seja desconhecida (MOSTAFAVI; MORRIS, 2012). Se formos capazes de prever corretamente essas interações proteicas, podemos inferir por exemplo, a função desconhecida de certas proteínas que interagem com outras proteínas cujas funções já são conhecidas (SHOEMAKER; PANCHENKO, 2007) (HOU, 2017). Portanto, o estudo das redes de interação proteína-proteína pode nos ajudar a compreender o funcionamento das proteínas dentro da célula.

Entretanto, não existem muitas alternativas computacionais disponíveis para realizar predições de redes de interação entre as proteínas de um organismo cuja sequência gênica acabou de ser elucidada.

1.1.2 Estado da Arte

Uma das principais ou talvez a principal alternativa computacional de predição das interações de proteínas até hoje, é o banco de dados de PPIs da ferramenta STRING (Ferramenta de Pesquisa para Recuperação de Genes em Interação) (SZKLARCZYK et al., 2019). STRING (<https://string-db.org/>) apresenta dados de anotação para mais de cinco mil genomas, espalhados por uma gama significativa de organismos. Características como vizinhança gênica conservada, perfil filogenético conservado, fusão de genes, características de ontologia genética (função molecular, processo e localização), coexpressão gênica, experimentos bioquímicos e evidências bibliográficas são conjugadas para criar uma força probabilística de crença de interação para pares de proteínas (SZKLARCZYK et al., 2019). O STRING realiza suas predições de PPIs usando uma técnica tradicional, a similaridade de sequência entre as proteínas cadastradas em seu banco de dados e as proteínas do genoma de consulta do usuário. Caso o genoma sob estudo não esteja presente no banco de dados do STRING, por ser uma cepa diferente de um organismo conhecido, ou ainda, por ser uma atualização de versão de genoma, então o usuário do STRING não poderá contar com uma predição para todos os pares de proteínas de seu genoma de consulta. Poderá contar apenas com uma predição para os pares de proteínas que possuem um par similar cadastrado no banco de dados do STRING. Acreditamos que o principal problema da predição de PPI com base na técnica de similaridade de sequência implementada no STRING, reside na novidade dos novos genomas. Sabe-se que pelo menos dez por cento das proteínas de cada novo genoma sequenciado e montado, não estão presentes em nenhum outro genoma conhecido (LAPIERRE; GOGARTEN, 2009). Isso significa que devemos esperar que em uma linhagem totalmente nova de *Escherichia coli* por exemplo, que possui em média 4 mil proteínas, exista pelo menos quatrocentas proteínas para as quais não se achará uma proteína similar no banco de dados do STRING. Isso resulta numa perda de aproximadamente $79.800 (400*(400-1)/2)$ possíveis pares de proteínas em interação. Ou ainda, 10% de possíveis interações negligenciadas em um genoma inédito, o que representa uma perda significativa na predição de PPIs.

Além das opções de predição através de similaridade de sequência frente a bancos de dados públicos como o STRING, muitos algoritmos de aprendizado de máquina tem sido aplicados ao problema de predição de interações proteicas obtendo percentuais satisfatórios de acurácia. Uma limitação das abordagens de aprendizado de máquina aplicadas a esse problema, decorre do fato desses algoritmos usarem padrões conhecidos para treinar um classificador. Como dito anteriormente sobre genomas inéditos, a cada novo genoma sequenciado e montado, 10% de seus genes serão exclusivos daquela sequência gênica espe-

cífica. Isso significa que não haverá um padrão de PPI previamente mapeado em qualquer modelo preditivo de aprendizado de máquina para esses genes inéditos, de modo que o preditor gerado provavelmente os classificaria como dados ruidosos.

Percebe-se que tanto através de ferramentas de bancos de dados públicos que fornecem informações sobre interações de proteínas, quanto por meio de abordagens de aprendizado de máquina, pelo menos 10% dos genes são negligenciados na predição de PPI para um genoma inédito. Mesmo não sendo genomas inéditos, existe a possibilidade desses genes não estarem cadastrados no banco de dados de uma ferramenta como o STRING. Ou ainda, de não haver similaridade entre algumas proteínas de consulta do usuário e as proteínas cadastradas no banco de dados utilizado. Isso impossibilitaria a predição de PPIs para essas proteínas. Tendo em vista a vasta variedade de proteínas em cada espécie, também existe a forte possibilidade de um modelo preditivo baseado em aprendizado de máquina, não ser generalista o suficiente para identificar interações confiáveis para as proteínas de todos os genomas bacterianos. Essas limitações das alternativas do estado da arte da predição computacional de interações proteicas, enfatizam a necessidade de criação de outras alternativas computacionais.

1.2 Objetivos e Desafios da Pesquisa

Nesta seção são apresentados os principais desafios deste trabalho, bem como os objetivos determinados em termos de objetivo geral e objetivos específicos.

1.2.1 Desafios

Realizar uma predição computacional *ab initio* das interações entre as proteínas de um dado organismo sob estudo, não é uma tarefa trivial. O grande desafio deste trabalho é desenvolver uma ferramenta computacional capaz de prever corretamente essas interações proteicas. O principal obstáculo é o tempo de processamento, pois esse método computacional depende de explorar bases de dados enormes de genomas. Isso é feito realizando inúmeras comparações par-a-par de proteínas em busca de evidências biológicas de interações proteicas para embasar as predições. Precisamos reunir um conjunto expressivo de genomas de referência para cada gênero bacteriano ou espécie envolvida em uma análise, a fim de realizar predições confiáveis (SNEL et al., 2000). Se utilizássemos as alternativas do estado da arte de algoritmos para comparação de sequências biológicas, o processamento computacional desses dados poderia durar dias tornando essa solução impraticável.

1.2.2 Objetivo Geral

Como objetivo geral e principal tem-se contribuir no campo da biologia computacional fornecendo um novo software autônomo de bioinformática para predição *ab initio* de redes de interação entre proteínas bacterianas. Com essa ferramenta almeja-se prever

redes de interações proteicas biologicamente corretas e confiáveis. Também é almejado realizar essa tarefa em tempo aceitável mesmo que utilizando um computador de configuração convencional. Como objetivo secundário foi definido a utilização desse software para analisar genomas de *Corynebacterium pseudotuberculosis*, *Corynebacterium diphtheriae* e *Rhodococcus*. Tais genomas são do grupo *Corynebacterium*, *Mycobacterium*, *Nocardia* e *Rhodococcus* (CMNR) de importância relevante para a medicina e economia. Também serão analisados alguns genomas inéditos do gênero bacteriano de *Dietzia* e *Aeromonas*, que foram recentemente sequenciados e montados pela Rede de Ciências Ômicas (RECOM) (<<https://www.recom-network.com/>>). Esses objetivos são alcançados progressivamente, na medida que cada objetivo específico listado a seguir, é atingido.

1.2.3 Objetivos Específicos

- ❑ Desenvolver um programa de computador para predição *ab initio* de redes de interação entre proteínas bacterianas;
- ❑ Utilizar o software desenvolvido para analisar genomas de diversos organismos de importância relevante para a medicina e economia;
- ❑ Validar biologicamente as predições de PPIs feitas pelo programa proposto;
- ❑ Comparar as redes de interação geradas pelo software desenvolvido com uma rede da principal ferramenta do estado da arte (STRING). O objetivo é estimar a qualidade das redes geradas tendo como referência uma rede do STRING.

1.3 Hipótese

Para suprir uma deficiência apontada na Subseção 1.1.2, concernente a atual indisponibilidade de ferramentas computacionais capazes de realizar predições de interações proteicas sem negligenciar proteínas inéditas, propõe-se um novo software autônomo de bioinformática denominado GenPPI. Esse programa realiza a predição de redes de interação proteica de modo *ab initio* (a partir dos dados contidos nos genomas). Como hipótese principal deste trabalho, admite-se a possibilidade de fazer predições de redes de PPI biologicamente corretas e confiáveis com a ferramenta proposta. Como hipótese secundária, assume-se que em decorrência de uma nova heurística para comparação de sequências de proteínas introduzida por este trabalho, seria capaz contornar o grande obstáculo da predição *ab initio*, o tempo de processamento. Como terceira e última hipótese, considera-se a possibilidade de gerar redes de interações proteicas com níveis de qualidade próximos da principal ferramenta do estado da arte (STRING). Tal qualidade é estimada em termos do percentual de interações proteicas em comum, preditas por cada ferramenta computacional para um mesmo organismo bacteriano selecionado. Além disso, também são verificadas características topológicas de redes complexas, capazes de fornecer uma

visão geral da qualidade de uma rede de PPI para análises biológicas diversas. Portanto, as perguntas associadas às hipóteses deste trabalho, são:

- ❑ A heurística proposta apresenta um nível de exatidão satisfatório na comparação par-a-par de sequências de aminoácidos de proteínas?
- ❑ O programa desenvolvido é capaz de fazer predições de interações proteicas biologicamente corretas?
- ❑ O GenPPI consegue realizar a predição *ab initio* em um tempo aceitável?
- ❑ As redes de interação preditas pela ferramenta proposta, apresentam qualidade satisfatória em comparação com uma rede do STRING?

1.4 Contribuições

Este trabalho de pesquisa contribui sob diversos aspectos que englobam desde a disponibilização de um novo software autônomo de bioinformática, até a geração de dados e informações inéditas para diversos organismos bacterianos. Para destacar as principais contribuições deste trabalho, menciona-se:

- ❑ Disponibilização de um novo programa autônomo de bioinformática para predição *ab initio* de redes de interação entre proteínas bacterianas.
- ❑ Contribuição com a rede de pesquisa RECOM na distinção de espécies do gênero bacteriano *Aeromonas*, recentemente sequenciadas e montadas pela rede. Alguns resultados e conclusões das análises feitas com esses genomas, foram incluídos em um artigo que, até o mês de Dezembro de 2020, ainda estava sendo produzido pelo grupo RECOM. É importante salientar que as análises computacionais feitas neste trabalho, se tornaram decisivas nas conclusões desse artigo sob redação. O motivo é que os métodos bioquímicos comumente utilizados para diferenciar bactérias não possuem exatidão suficiente para divisar entre espécies do gênero bacteriano *Aeromonas*.
- ❑ Uma nova heurística para comparação par-a-par de sequências de aminoácidos de proteínas, com maior exatidão e menor tempo de processamento que o principal algoritmo heurístico do estado da arte (BLASTp).

1.5 Organização da Dissertação

No Capítulo 2 são apresentados os conceitos teóricos fundamentais deste trabalho e também alguns trabalhos correlatos que ditam o estado da arte. Na sequência, a ferramenta proposta e sua implementação são descritas com detalhes no Capítulo 3. Os

experimentos e análises de resultados que comprovam as hipóteses levantadas, são apresentados no Capítulo 4. Finalmente, o Capítulo 5 apresenta a conclusão do trabalho, bem como as principais contribuições e os trabalhos futuros.

Fundamentação Teórica

Neste capítulo os conceitos teóricos que fundamentam este trabalho, são divididos em dois aspectos: fundamentação biológica e fundamentação matemática/computacional. Por final, são apresentados alguns trabalhos correlatos do estado da arte.

2.1 Fundamentação Biológica

Nesta seção é apresentada a fundamentação biológica ligada ao trabalho.

2.1.1 Genomas

Genoma é a sequência completa de DNA (ácido desoxirribonucleico) de um organismo, ou seja, o conjunto de todos os genes de um ser vivo. O DNA codifica todas as informações necessárias para os processos de células individuais e, conseqüentemente, as funções e características herdadas dos organismos. O DNA de uma célula compreende o genoma dessa célula. Um genoma é toda a informação hereditária codificada no DNA de um organismo (NG; KHOR, 2017).

2.1.2 Proteínas

As proteínas são longas cadeias poliméricas que estão presentes em todas as células vivas e possuem funções biológicas específicas. As proteínas são constituídas de aminoácidos dentre os quais existem 20 principais tipos¹ Os aminoácidos têm propriedades químicas distintas (por exemplo, a carga elétrica de suas cadeias laterais). Os aminoácidos podem se ligar a outros por ligações covalentes para formar polipeptídeos. As proteínas consistem em pelo menos um polipeptídeo, ou seja, um composto orgânico constituído por uma sequência de até 20 tipos de aminoácidos conectados por meio de ligações peptídicas. As proteínas são responsáveis por uma variedade impressionante de funções, são componentes estruturais fundamentais das células e também estão envolvidas em quase todas as

¹ Apenas vinte deles estão presentes no chamado "código genético padrão", porém outros seis tipos podem ser sintetizados em circunstâncias especiais (AMBROGELLY; PALIOURA; SÖLL, 2007).

funções celulares, como transporte, regulação hormonal, metabolismo, respiração, reparo e controle de genes (KESKIN et al., 2008).

2.1.3 Do DNA às Proteínas

O DNA consiste em duas longas cadeias de nucleotídeos, sendo que cada nucleotídeo é composto de uma molécula de açúcar, uma molécula de fosfato e uma base nitrogenada. Todos os organismos vivos são codificados por quatro nucleotídeos: adenina (A), timina (T), guanina (G) e citosina (C). A ordem particular dos nucleotídeos em qualquer uma das cadeias de DNA é chamada de sequência de DNA. O DNA possui uma estrutura de dupla hélice, composta por moléculas de açúcar, grupo fosfato e bases (A, G, C, T). As duas fitas de DNA são complementares, o que significa que elas contêm a mesma informação genética (a informação é duplicada) e são mantidas juntas por ligações fracas de hidrogênio. Na sequência de DNA contém instruções para a síntese de cada proteína. Essas são as seções específicas da sequência de DNA geralmente chamadas de genes (NG; KHOR, 2017).

A maneira como as informações armazenadas no DNA são transmitidas para a formação das proteínas é chamada de dogma central da biologia molecular (CRICK, 1970). Um esquema simplificado desse processo está ilustrado na Figura 2.

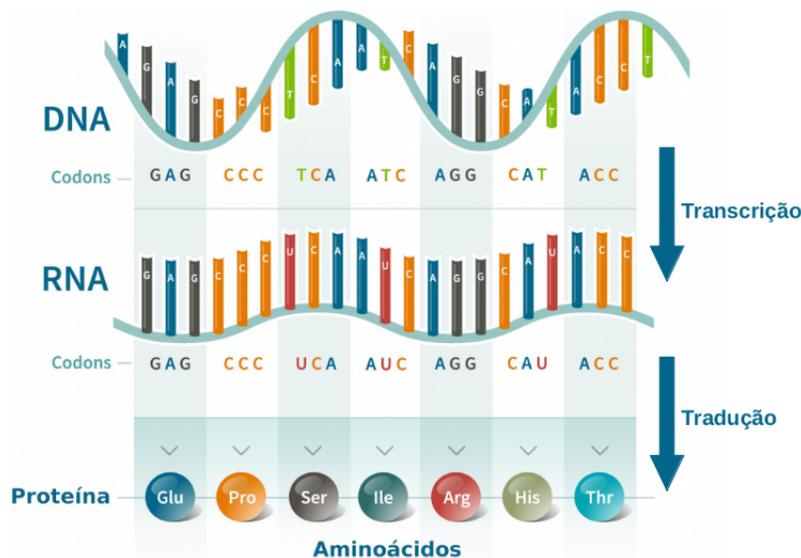


Figura 2 – Dogma central da biologia molecular

Fonte: Adaptado de (ANCESTRYDNA, 2020)

Esse processo é representado por duas etapas principais como a seguir:

- Transcrição (DNA => RNA): é o processo em que as informações codificadas em um segmento específico da sequência de DNA (ou gene) são passadas para uma molécula

de RNA chamada RNA mensageiro (mRNA). Moléculas de mRNA são semelhantes ao DNA. Elas também são uma cadeia de nucleotídeos, mas contêm apenas uma fita e usam diferentes bases nitrogenadas e açúcares. Além disso, o mRNA é menor porque contém as informações relacionadas a apenas um gene. O processo pelo qual os genes são transcritos para uma molécula de mRNA é geralmente chamado de expressão gênica (CRICK, 1970).

- Tradução (mRNA => Proteína): é o processo em que a informação genética agora codificada no mRNA é usada para sintetizar uma proteína específica. Esse processo é mediado por outras macromoléculas chamadas ribossomos e também por outros tipos de moléculas de RNA. A informação genética é traduzida de uma cadeia de nucleotídeos do mRNA para uma cadeia de aminoácidos. Isso é feito usando o código genético, onde um trigêmeo de nucleotídeos (códon) está associado a um aminoácido específico, conforme ilustrado pela Figura 2. Há um total de 26 aminoácidos diferentes, mas apenas 20 deles estão presentes no chamado código genético padrão (AMBROGELLY; PALIOURA; SÖLL, 2007). A sequência final de aminoácidos gerados corresponde ao que conhecemos como proteína (CRICK, 1970).

Os 20 aminoácidos presentes no código genético padrão, suas abreviações e siglas, são mostrados na Tabela 1.

Tabela 1 – Tabela dos 20 aminoácidos.

Aminoácido	Abreviação	Sigla
Glicina	Gly	G
Alanina	Ala	A
Cisteína	Cys	C
Serina	Ser	S
Treonina	Thr	T
Valina	Val	V
Asparagina	Asn	N
Aspartato	Asp	D
Fenilalanina	Phe	F
Histidina	His	H
Isoleucina	Ile	I
Leucina	Leu	L
Prolina	Pro	P
Tirosina	Tyr	Y
Triptofano	Trp	W
Glutamato	Glu	E
Glutamina	Gln	Q
Metionina	Met	M
Arginina	Arg	R
Lisina	Lys	K

2.1.4 Estrutura das Proteínas

As proteínas são polímeros constituídos por cadeias de aminoácidos. A estrutura e a forma das proteínas (como a cadeia de aminoácidos se dobra num espaço tridimensional) é relevante para determinar sua função específica. A estrutura da proteína pode ser descrita em vários níveis. O primeiro nível é denominado estrutura primária e corresponde à sequência linear de aminoácidos. A estrutura secundária se refere a como a cadeia de aminoácidos da proteína é organizada em um espaço tridimensional, formando ligações de hidrogênio com ela mesma. Existem dois componentes principais na estrutura secundária: hélices alfa e folhas beta. A estrutura terciária é produzida quando elementos da estrutura secundária se dobram entre eles. finalmente, a estrutura quaternária está relacionada ao arranjo espacial da proteína. A Figura 3 apresenta uma representação esquemática dessas conformações estruturais. A estrutura final de uma proteína determina a sua função (BONETTI, 2012).

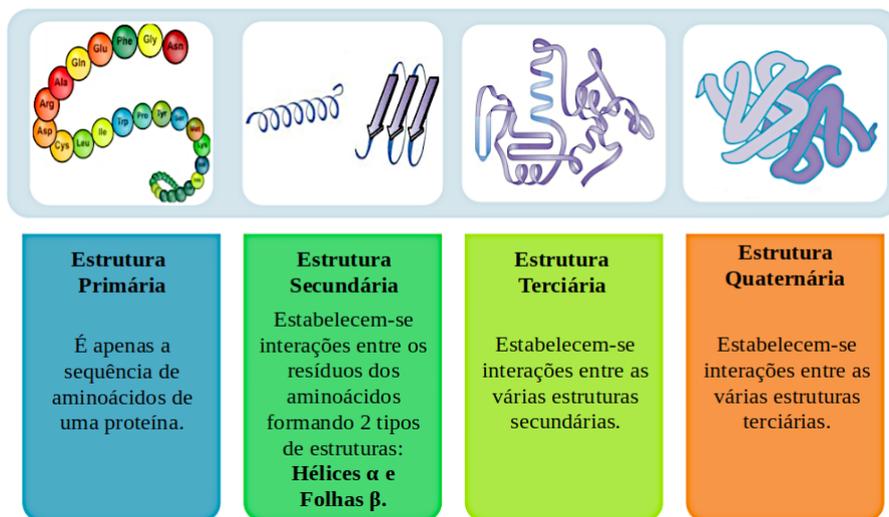


Figura 3 – As quatro estruturas de uma proteína

2.1.5 Função das Proteínas

As proteínas estão envolvidas em quase todas as funções desempenhadas em uma célula. Entre essas, encontramos (KESKIN et al., 2008):

- ❑ Enzimas que catalisam por exemplo, muitas das reações do metabolismo;
- ❑ Proteínas estruturais, como o colágeno que é a principal proteína do tecido conjuntivo em animais;
- ❑ Proteínas regulatórias, como fatores de transcrição que regulam a transcrição de genes;

- Moléculas de sinalização, como certos hormônios, como a insulina e seus receptores;
- Proteínas de defesa, como anticorpos do sistema imunológico.

Atualmente, devido à disponibilidade de técnicas de sequenciamento de alto rendimento, conhecemos a sequência completa do genoma (DNA) de várias espécies. Com isso, também somos capazes de obter a sequência de aminoácidos da maioria das proteínas. Entretanto, a função de grande parte dessas proteínas permanece desconhecida. Portanto, a predição das funções das proteínas ainda é uma das áreas de pesquisa mais importantes em bioinformática.

O estudo das interações entre as proteínas pode ajudar nessa tarefa. Se formos capazes de prever corretamente essas interações, poderíamos inferir por exemplo, a função desconhecida de certas proteínas que interagem com outras proteínas que possuem funções já conhecidas (SHOEMAKER; PANCHENKO, 2007). Assim, o estudo das interações proteicas de um organismo pode nos ajudar a entender como as proteínas funcionam dentro da célula.

2.1.6 Interação Proteína-Proteína

Como já foi mencionado anteriormente, muitas proteínas não funcionam sozinhas, mas em coordenação com outras proteínas. Isso gera interações de pares de proteínas e complexos proteicos. As interações proteína-proteína (PPIs) participam de todos os processos biológicos importantes em organismos vivos, como catalisar reações metabólicas, replicação de DNA, transcrição de DNA, responder a estímulos e transportar moléculas de um local para outro (PENG et al., 2017).

As interações entre as proteínas podem ser divididas em dois principais tipos (BROWNE et al., 2010):

1. Interações diretas: envolvem contato físico direto entre as proteínas.
2. Associação funcional: o par de proteínas em interação não tem um contato físico direto, mas indiretamente interage para desempenhar funções específicas na célula.

As proteínas exercem suas funções biológicas participando de uma rede complexa de interações. O interesse da comunidade biomédica na reconstrução completa de redes de interação entre proteínas para várias espécies, vem crescendo progressivamente. Esse interatoma (conjunto de todas as interações moleculares em uma determinada célula), termo usado por (SANCHEZ et al., 1999), proporciona uma grande percepção do comportamento das proteínas dentro da célula e, posteriormente, permitiu pesquisas mais focadas. Por exemplo, redes de interação entre proteínas de espécies bacterianas podem evidenciar os mecanismos pelos quais um patógeno coloniza e infecta seu hospedeiro. A infecção por esses mecanismos poderia ser evitada, pois as interações responsáveis pela infecção

no hospedeiro poderiam ser identificadas, e medicamentos seriam projetados para atingir e neutralizar essas interações específicas (ARKIN; WELLS, 2004). Um trabalho recente nesse sentido pode ser visto em (ZHANG et al., 2020).

2.1.7 Pan-genoma

Nos campos da biologia molecular e da genética, um pan-genoma (ou supragenoma) é o conjunto completo de genes para todas as cepas (variantes genéticas ou linhagens de um organismo) dentro de um clado (grupo de organismos que descendem de um ancestral comum). O pan-genoma inclui: o genoma do núcleo contendo genes presentes em todas as cepas do clado, o genoma acessório contendo genes presentes em um subconjunto dessas cepas, e genes específicos de cada cepa. (VERNIKOS et al., 2015). Uma ilustração de um pan-genoma é dada pela Figura 4.

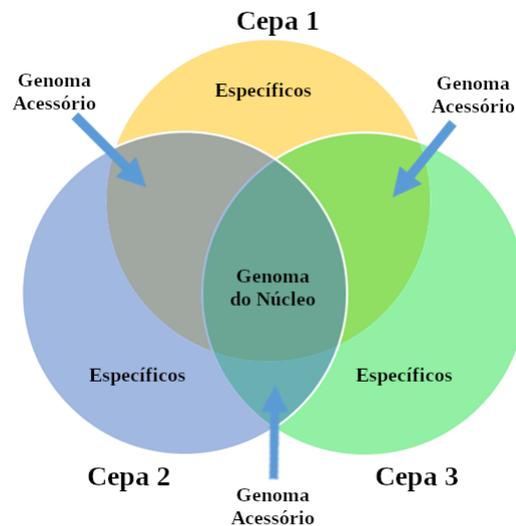


Figura 4 – Ilustração do pan-genoma de três cepas fictícias

Fonte: Adaptado de (WHEATGENOMES, 2020)

2.2 Fundamentação Matemática e Computacional

Nesta seção é apresentada a fundamentação matemática e computacional ligada a este trabalho. Isso inclui uma revisão literária dos principais algoritmos de comparação de sequências biológicas, e das principais alternativas para predição computacional de interações entre proteínas.

2.2.1 Grafos

As interações proteicas de um organismo são modeladas por meio de redes complexas. Uma rede complexa é uma representação gráfica de um conjunto de elementos físicos

diferentes (ou estruturas abstratas) representadas por nós, interligados entre si por meio de links, segundo uma determinada regra de conexão formando uma certa estrutura topológica. De acordo com (CHEN; WANG; LI, 2014), para que seja possível descrever as características comuns e as propriedades inerentes a diferentes tipos de redes complexas, faz-se necessária uma ferramenta eficiente e igualmente rigorosa de análise, oferecida pela matemática no campo da Teoria dos Grafos.

Um grafo é uma estrutura matemática frequentemente usada para definir um conjunto de objetos e suas relações. Um grafo é definido formalmente como um par ordenado $G = (V, E)$, em que V é um conjunto finito, não vazio de vértices (ou nós), e E é o conjunto das arestas (ou links) existentes entre os vértices de V . A quantidade de vértices de G é dada por $n = |V|$ e a quantidade de arestas, por $m = |E|$ (NEWMAN, 2018).

Na sequência são dadas algumas definições básicas e propriedades fundamentais sobre grafos para este trabalho:

- Dois vértices ligados por uma aresta são ditos **adjacentes**.

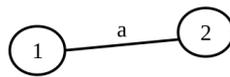


Figura 5 – Vértices adjacentes

- Uma aresta que ligue dois vértices é dita **incidente** de cada um dos vértices, portanto a aresta ‘a’ no grafo da Figura 5 é incidente dos vértices 1 e 2.
- O número de arestas incidentes num vértice, é chamado de **grau desse vértice**. O **grau máximo** do grafo é o maior dos graus dos vértices ($\Delta(G)$), conseqüentemente o **grau mínimo** ($\delta(G)$) é o menor dos graus dos vértices.

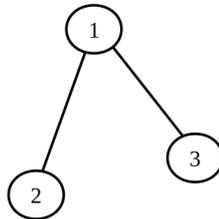


Figura 6 – Grau dos vértices. O vértice 1 tem duas arestas incidentes, portanto grau 2. Os vértices 2 e 3 tem uma aresta incidente, portanto, ambos tem grau 1. Assim, o grau máximo desse grafo é 2 e o grau mínimo é 1.

- O **grau médio** de um grafo é calculado como a média aritmética dos graus de todos os seus vértices.

- As arestas podem ser **direcionadas** ou **não direcionadas**. Se uma aresta que liga os vértices A e B não for direcionada, significa que A está vinculado a B e B está vinculado a A. Em vez disso, se a aresta que liga os vértices A e B for direcionada de A para B, isso significa que existe apenas o vínculo de A em B. A Figura 7 ilustra a diferença entre um grafo não direcionado e um grafo direcionado.

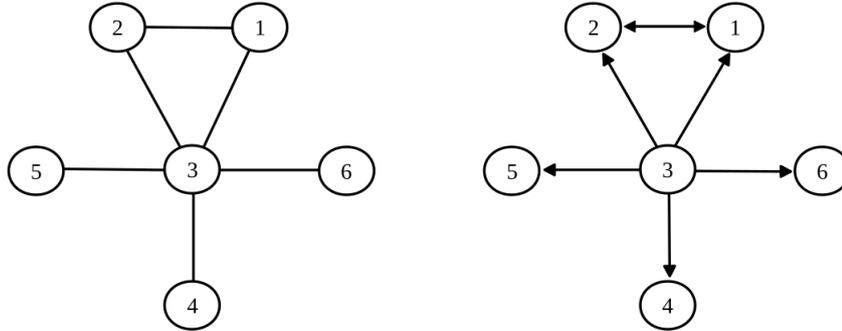


Figura 7 – A diferença entre dois grafos com base em sua direcionalidade. A figura à esquerda mostra um grafo não direcionado, enquanto a figura à direita mostra um grafo direcionado.

- As arestas podem possuir pesos que denotam relações por uma grandeza específica entre os vértices caracterizando assim, um grafo ponderado. Em um **grafo ponderado**, um valor numérico é atribuído a cada aresta representando grandezas específicas, tais como distâncias, altitudes, capacidades, fluxos, força de interação entre elementos, etc. A Figura 8 apresenta um exemplo de grafo ponderado.

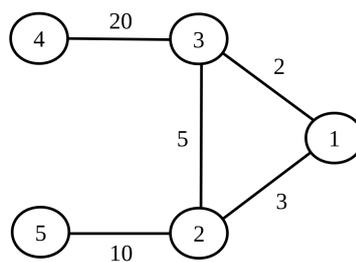


Figura 8 – Exemplo de um grafo ponderado

- Um **grafo denso** é um grafo em que o número de arestas é próximo ao número máximo possível de arestas. O oposto, um grafo com apenas algumas arestas, é um **grafo esparso**. A **densidade** de um grafo é representada pela razão entre o número de arestas do grafo e o número máximo possível de arestas nesse grafo. No contexto de redes de interação entre proteínas, baixos valores de densidade representam redes mais adequadas para análises biológicas diversas. Isso porque altas densidades

representam redes altamente conectadas, nas quais não se consegue inferir muitas informações biológicas (KOUTROULI et al., 2020).

- Um **grafo completo** ou **totalmente conexo** é um grafo simples e não direcionado que possui a característica de que todo vértice do grafo é adjacente aos demais. Ou seja, cada par de vértices distintos é conectado por uma aresta única. A quantidade de arestas em um grafo completo é igual a $n(n - 1)/2$, onde n é o número de vértices do grafo (KOUTROULI et al., 2020). A Figura 9 apresenta um exemplo de grafo completo com 6 vértices e 15 arestas.

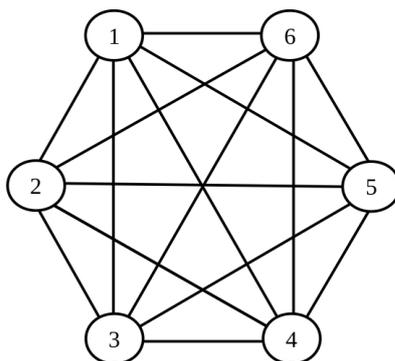


Figura 9 – Exemplo de um grafo completo ou totalmente conexo

As redes de interações de proteínas são redes complexas e podem ser convenientemente modeladas com grafos não direcionados e ponderados. Os vértices representam as proteínas, as arestas indicam as interações proteicas existentes, e o peso das arestas representa uma estimativa de confiabilidade na predição de interação para um determinado par de proteínas.

As redes de PPI tem a maioria dos vértices apresentando poucas arestas, contrastando com a existência de alguns vértices que apresentam um elevado número de arestas. Portanto, a predição de PPI nunca produzirá um grafo completo, ou seja, seu número de arestas sempre será menor que $n(n - 1)/2$, onde n representa o número de vértices (proteínas) do grafo.

2.2.2 Common Lisp

Em 1956, John McCarthy especificou o Lisp (o nome deriva de "List Processing"), a segunda mais antiga linguagem de programação de alto nível atualmente em uso generalizado (somente o Fortran é mais velho, em um ano). McCarthy era um pesquisador de inteligência artificial, e muitos dos recursos que ele construiu na versão inicial da linguagem Lisp a tornaram uma excelente linguagem para a programação de IA. Assim como a linguagem Prolog, Lisp é adequada para problemas de inteligência artificial principalmente por facilitar o processamento de símbolos. Durante a explosão da IA na década de

1980, o Lisp permaneceu como uma ferramenta favorita dos programadores que escreviam software para resolver problemas difíceis, como prova automatizada de teoremas e visão computacional (SEIBEL, 2006).

Lisp introduziu a ideia do loop de leitura-avaliação-impressão (REPL – read-eval-print loop). Esse ambiente lê uma linha de código Lisp, avalia, imprime o resultado e em seguida retorna ao início do processo. Essa mudança permitiu o desenvolvimento incremental em que o desenvolvedor gradualmente escreve uma função, tenta chamá-la, testa se funciona e depois passa para o próximo trecho de código. Esse sistema tornou o desenvolvimento de software significativamente mais rápido e também mais fácil de lidar com bugs, pois o desenvolvedor pode testar cada parte do programa por vez e identificar mais facilmente onde fez algo errado no código-fonte (CHISNALL, 2009).

De acordo com (HOYTE, 2008), a maior vantagem do Lisp como linguagem de programação é seu sistema de macros. Com elas, o desenvolvedor pode fazer coisas que simplesmente não podem ser feitas em outras linguagens de programação. As macros permitem introduzir novas construções de linguagem e sintaxe, descrevendo-as no próprio Lisp. Isso significa que uma nova construção de sintaxe pode ser adicionada como uma biblioteca, sem precisar alterar o compilador. Por exemplo, para estender o Python ou Java com um novo tipo de loop, seria preciso estender a sintaxe e a semântica da linguagem e, em seguida, implementar esse novo recurso no compilador/interpretador. Em Lisp bastaria escrever uma macro e a mesma seria usada pelo compilador para gerar o código nativo a ser compilado. Isso permitiu que o desenvolvedor crie abstrações sobre a linguagem central e a biblioteca padrão que expressem mais claramente o que se deseja implementar. Assim, os programadores que ganham experiência com as macros do Lisp, descobrem que com elas é possível realizar feitos incríveis de abstração, produtividade, eficiência e segurança de código.

O Lisp mudou muito desde seus primeiros dias e vários dialetos foram criados ao longo de sua história. Hoje, os dialetos Lisp de uso geral mais conhecidos são, Common Lisp, Scheme e Clojure. Common Lisp é uma linguagem multi-paradigma que suporta programação procedimental, funcional e orientada a objetos. Além de ser adequada para problemas de inteligência artificial por suas facilidades de processamento simbólico, o Common Lisp é uma das linguagens mais bem-sucedidas no mundo após a linguagem C. Possibilita a construção de programas complexos com uma pequena fração de tempo e código quando comparada com a linguagem C, sem deixar a desejar em termos de velocidade de execução (NORVIG, 1992).

O Common Lisp fornece um nível de flexibilidade ausente em linguagens como Java, Python e C++. Apresenta facilidades poderosas para executar programação orientada a objetos e vários recursos de programação não existentes em outras linguagens. Esses fatores do Common Lisp somados ao seu sistema exclusivo de macros já mencionado, possibilita a construção de programas complexos com um esforço menor.

2.2.3 Heurística

Etimologicamente a palavra heurística vem da palavra grega *Heuriskein* que significa descobrir e que deu origem ao termo *Eureca*. Em ciência da computação, inteligência artificial e otimização matemática, uma heurística é uma técnica projetada para resolver um problema mais rapidamente quando os métodos clássicos são muito lentos. Uma heurística também pode ser utilizada visando encontrar uma solução aproximada quando os métodos clássicos falham em encontrar uma solução exata (PEARL, 1984). De certa forma, pode ser considerado um atalho.

O objetivo de uma heurística é produzir uma solução em um prazo razoável, que seja boa o suficiente para resolver o problema em questão. Essa solução pode não ser a melhor de todas as soluções para esse problema ou pode simplesmente se aproximar da solução exata. Mas ainda é valiosa porque encontrá-la não requer um tempo proibitivo. Uma programação heurística emprega um método prático não garantido que seja ideal ou perfeito, mas suficiente para atingir um objetivo imediato (PEARL, 1984).

As heurísticas estão subjacentes a todo o campo da inteligência artificial sendo utilizadas em situações em que não existem algoritmos conhecidos.

Os critérios para decidir se deve-se usar uma heurística para resolver um determinado problema, incluem:

- ❑ **Otimalidade:** quando existem várias soluções para um determinado problema, a heurística garante que a melhor solução será encontrada? É realmente necessário encontrar a melhor solução?
- ❑ **Completude:** quando existem várias soluções para um determinado problema, a heurística pode encontrar todas? Nós realmente precisamos de todas as soluções? Muitas heurísticas destinam-se apenas a encontrar uma solução.
- ❑ **Exatidão:** a heurística pode fornecer um grau de confiança satisfatório para a solução pretendida? A margem de erro na solução é excessivamente grande?
- ❑ **Tempo de execução:** tempo de processamento da heurística proposta. Há heurísticas que convergem mais rapidamente que outras. Algumas são apenas marginalmente mais rápidas que os métodos exatos.

A pesquisa por heurísticas é uma pesquisa realizada por meio da quantificação de proximidade a um determinado objetivo. Diz-se que se tem uma boa heurística se o resultado produzido por ela estiver muito próximo do objetivo; diz-se de má heurística se o resultado estiver muito longe do objetivo (PEARL, 1984).

2.2.4 Revisão Literária dos Principais Algoritmos Para Comparação de Sequências Biológicas

Em Bioinformática uma das operações mais básicas e relevantes é a comparação de sequências biológicas. É amplamente realizada para determinar o nível de similaridade entre sequências e inferir características comuns entre espécies (DURBIN et al., 1998). Essa análise é de extrema importância na predição de redes de PPI baseada no contexto genômico, pois possibilita a detecção de evidências de eventos evolutivos que indicam interações entre proteínas. O resultado de uma operação de comparação de duas sequências, é um percentual de identidade entre elas (score).

A comparação de sequências biológicas geralmente é feita através do alinhamento de sequências, que é definido como um pareamento entre os caracteres das sequências. O alinhamento ótimo de sequências procura encontrar, entre todas as possibilidades de pareamento de duas sequências, um alinhamento que produza o maior score de identidade possível de acordo com o tipo de alinhamento. Dentre os principais tipos de alinhamento, podemos citar o alinhamento global que considera todos os caracteres de ambas as sequências, e o alinhamento local que considera apenas substrings das sequências. É importante salientar que o alinhamento local não se preocupa em realizar alinhamentos ótimos. Aqui, são apresentados os algoritmos Needleman-Wunsch e BLAST que são, respectivamente, as principais alternativas para alinhamento global e local de sequências biológicas até hoje.

2.2.4.1 Algoritmo Needleman-Wunsch

O algoritmo Needleman-Wunsch (NEEDLEMAN; WUNSCH, 1970) foi proposto por Saul Needleman e Christian Wunsch na década de 1970. Trata-se de um algoritmo para alinhamento global par-a-par de sequências biológicas. Isso significa que ele alinha pares de sequências em todo o seu comprimento. É importante ainda destacar que ele é ótimo, ou seja, retorna o alinhamento de maior percentual de identidade de sequência.

O algoritmo Needleman-Wunsch resolve o problema de alinhamento ótimo através da técnica de programação dinâmica com uma complexidade computacional igual a $\mathcal{O}(mn)$, onde m e n representam o tamanho das sequências a serem comparadas. A técnica de programação dinâmica consiste em resolver uma instância do problema a partir de computações já realizadas para instâncias menores do mesmo problema (SINGH, 2015).

O gargalo mais importante desse algoritmo é a complexidade temporal que é quadrática. Seu uso de memória também é um fator limitante, pois a complexidade quadrática de espaço $\mathcal{O}(mn)$, restringe bastante o tamanho das sequências a serem comparadas. Isso fica evidente quando comparamos sequências muito longas. Sendo assim, esse algoritmo demanda um alto poder de processamento e uma grande quantidade de memória. Por esse motivo, o uso de algoritmos exatos como o Needleman-Wunsch, foi considerado inviável por muito tempo. Com isso, algoritmos heurísticos surgiram para acelerar o procedimento

de comparação de sequências, embora sem garantir a produção de um resultado ótimo.

2.2.4.2 BLAST

O BLAST (ferramenta básica de busca de alinhamento local) (ALTSCHUL et al., 1990) é o algoritmo mais utilizado na área da bioinformática para comparar sequências biológicas como por exemplo, sequências de aminoácidos de proteínas ou nucleotídeos de sequências de DNA. A ferramenta BLAST é uma ferramenta muito utilizada dentro da bioinformática por ser uma ferramenta muito rápida quando comparada com algoritmos exatos como o Needleman-Wunsch, e também por existir variações do BLAST para comparação de sequências biológicas diversas.

Uma pesquisa BLAST permite comparar uma sequência fornecida em uma consulta com uma biblioteca ou base de dados de sequências, e identificar sequências biológicas que se assemelham à sequência de consulta. O BLAST localiza regiões de similaridade local entre sequências. O programa compara sequências de nucleotídeos ou de proteínas e calcula a significância estatística das correspondências.

Com o algoritmo do BLAST consegue-se aumentar o desempenho em termos de velocidade por se diminuir o espaço de busca ou número de comparações que são feitas. No caso dos algoritmos ótimos cada nucleotídeo/aminoácido é comparado com todos da outra sequência gerando uma complexidade computacional quadrática, o que não acontece no BLAST. O BLAST implementa um algoritmo heurístico para comparação de sequências biológicas que é muito mais rápido que algoritmos exatos como o Needleman-Wunsch. Embora o BLAST seja mais rápido que o Needleman-Wunsh, ele não garante alinhamentos ótimos. Essa ênfase na velocidade é de vital importância para tornar a solução praticável frente aos enormes bancos de dados genômicos atualmente disponíveis.

2.2.5 Revisão Literária de Ferramentas e Métodos Computacionais Para Predição de PPI

O objetivo desta subseção é fornecer uma revisão sobre a predição de interações entre proteínas através das alternativas computacionais do estado da arte. Inicialmente são apresentados os principais bancos de dados públicos que fornecem informações sobre interações de proteínas conhecidas. Em seguida são descritos os métodos principais de predição *ab initio* de PPI, baseados em evidências biológicas contidas nos genomas. Por final, abordagens de aprendizado de máquina são citadas e comentadas em geral. Aqui, o termo **gene** tem o mesmo significado semântico do termo **proteína**.

2.2.5.1 Bancos de Dados Públicos

Vários projetos internacionais anteriores de genômica acumularam uma grande quantidade de informações sobre interações de proteínas, que estão disponíveis em vários bancos

de dados públicos. Com o aprofundamento dessa área de pesquisa, esses bancos de dados se tornaram gradualmente um recurso importante para comunidade científica da área. A predição de PPIs baseada em similaridade de sequência contra bancos de dados de interações entre proteínas, compara sequências biológicas de consulta contra as sequências depositadas nesses bancos de dados para inferir interações. O conhecimento sobre as interações proteicas presente nesses bancos de dados é provenientes de toda sorte de técnicas para predição de PPIs. Tais como vizinhança gênica conservada, perfil filogenético conservado, fusão de genes, características de ontologia genética (função molecular, processo e localização), coexpressão gênica, experimentos bioquímicos e evidências bibliográficas.

Existem muitos bancos de dados diferentes que fornecem informações sobre interações de proteínas. Entre os mais populares, estão: STRING (Search Tool for Recurring Instances of Neighbouring Genes) (SZKLARCZYK et al., 2019); BioGRID (Biological General Repository for Interaction Datasets) (STARK et al., 2006); DAVID (Database for Annotation, Visualization and Integrated Discovery) (JIAO et al., 2012); HPRD (Human Protein Reference Database) (PRASAD et al., 2009); HIPPIE (Human Integrated Protein-Protein Interaction rEference) (CHEN; PANDEY; NGUYEN, 2017); KEGG (Kyoto Encyclopedia of Genes and Genomes) (KANEHISA et al., 2019); e Metascape (A Gene Annotation and Analysis Resource) (ZHOU et al., 2019). Como mostra a Tabela 2, os bancos de dados variam no número de genes e de pares de proteínas em interação.

Tabela 2 – Número de genes e interações conhecidas disponíveis para consulta em bancos de dados. Os dados fornecidos nesta tabela foram colhidos nos sites desses banco de dados em Outubro de 2020.

Bando de Dados	Versão	Número de Genes	Número de Interações
BioGRID	3.5.185	78.567	1.415.932
DAVID	6.8	Não informado	Não informado
HPRD	Release 9	30.047	41.327
HIPPIE	2.2	16.792	340.629
KEGG	95.2	32.569.374	20.411.659.773
Metascape	3.5	Não informado	Não informado
STRING	11.0	24.584.628	3.123.056.667

Ferramentas tão notáveis não precisam de maiores apresentações possuindo úteis análises de enriquecimento e interfaces amigáveis para biólogos. Muitas dessas bases de dados permitem exportar relatórios e continuar estudos adicionais usando softwares necessários como Python (ROSSUM; DRAKE, 2009), Cytoscape (LOPES et al., 2010), R (TEAM et al., 2013), UALCAN (CHANDRASHEKAR et al., 2017), MCODE (BADER; HOGUE, 2003) e GEPHI (LEONARD; GRAHAM; BONACUM, 2004). Deve-se também mencionar o número quase infinito de bibliotecas existentes que são implantadas anualmente em todos esses softwares. Por exemplo, tais bibliotecas nos permitem focar em genes candidatos, genes diferencialmente expressos (DEGs), estrutura terciária de interações de proteínas e muitos outros recursos úteis.

Um pesquisador que lida com organismos modelos não enfrentará problemas na obtenção de percepções sobre interações de proteínas utilizando todos os bancos de dados mencionados anteriormente. Por exemplo, ao estudar *H. sapiens*, *M. musculus*, *R. norvegicus*, *D. rerio*, *D. melanogaster*, *C. elegans*, *S. cerevisiae*, *A. thaliana*, *S. Pombe* e *P. falciparum*. Se estiver usando os bancos de dados STRING (SZKLARCZYK et al., 2019), Metascape (ZHOU et al., 2019) e DAVID (JIAO et al., 2012), uma lista de genes de entrada é suficiente para que esses banco de dados retornem interações confiáveis. No entanto, ao lidar com organismos inéditos, um pesquisador não poderá contar com uma predição de PPI confiável. O motivo é que essas alternativas computacionais aplicam uma técnica tradicional para realizar uma predição, a similaridade de sequência entre as proteínas cadastradas em seu banco de dados e as proteínas do genoma de consulta do usuário. Caso o genoma sob estudo não esteja presente nesses bancos de dados, por ser uma cepa diferente de um organismo conhecido, ou ainda, por ser um genoma totalmente inédito, então um pesquisador não poderá contar com uma predição para todos os pares de proteínas de seu genoma de consulta. Apenas para os pares de proteínas que possuem um par similar cadastrado no banco de dados em questão. Esse cenário é mais provável de acontecer ao estudar procariotos. O estudo dos procariotos traz à luz dezenas de novos genomas e milhares de genes inéditos constantemente.

Como já foi mencionado anteriormente, é fato que pelo menos 10% dos genes de genomas inéditos, são específicos daquela espécie, não existindo em nenhum outro genoma conhecido (LAPIERRE; GOGARTEN, 2009). Assim, esses 10% de genes inéditos são negligenciados na predição de PPI através dos bancos de dados públicos que fornecem informações sobre interações conhecidas. Além disso, acreditamos que os bancos de dados públicos não acompanharão a grande e crescente quantidade de genomas sequenciados e montados constantemente. Esses fatores enfatizam a necessidade de novas alternativas para predição de interações entre proteínas.

2.2.5.2 Predição *Ab Initio* de PPIs Baseada em Evidências Biológicas no Contexto Genômico

A predição *ab initio* de PPIs baseada em evidências biológicas no contexto genômico, é feita a partir das sequências gênicas dos organismos sob análise. Genomas são analisados em conjunto visando encontrar evidências de eventos evolutivos que indicam interações entre proteínas (HIROSE, 2012). Como dito anteriormente, muitas proteínas exercem suas funções biológicas interagindo umas com as outras (RAMANATHAN; PORTER; KHAVARI, 2019). As proteínas interagentes necessárias para realizar funções específicas em um organismo, sofrem certas pressões seletivas durante a evolução das espécies (HIROSE, 2012). Por exemplo, uma pressão para se manterem próximas umas das outras na fita de DNA, a fim facilitar a geração de seus produtos proteicos. Isso influencia na estrutura dos genomas deixando rastros que podem ser detectados ao analisar vários genomas em con-

junto (HIROSE, 2012). Esses rastros (evidências) são decorrentes de eventos evolutivos comumente utilizados para inferir computacionalmente interações entre proteínas. Nesta subseção são apresentados três tipos desses eventos: (i) vizinhança gênica conservada, (ii) fusão gênica e (iii) perfil filogenético conservado. Além desses três, existem outros que também podem ser utilizados na predição computacional de PPIs (VALENCIA; PAZOS, 2002). Esse método requer um esquema de pontuação que integre os tipos de eventos evolutivos analisados, e dê um valor de confiança para cada interação inferida (MERING et al., 2003). Essa abordagem não visa a predição de interações físicas entre as proteínas, mas sim de proteínas funcionalmente associadas. Porém, proteínas funcionalmente associadas também podem interagir fisicamente entre si (HIROSE, 2012).

Uma vantagem dessa abordagem *ab initio* em relação aos bancos de dados públicos, é que a mesma não negligencia as proteínas inéditas dos novos genomas. O motivo é que nessa abordagem, a predição de PPIs é feita a partir das sequências gênicas dos próprios organismos de interesse. Assim, todas as proteínas dos genomas selecionados são analisadas e nenhuma é ignorada. Outra vantagem é que devido ao número crescente de genomas sequenciados e montados, existe agora uma grande quantidade de sequências gênicas disponíveis e passíveis de serem utilizadas nesse método.

A seguir é percorrido sobre a predição de PPIs baseada nos eventos evolutivos de vizinhança gênica conservada, fusão gênica e perfil filogenético conservado.

Eventos de **Vizinhança Gênica Conservada** (do inglês, Conserved Neighborhood (CN)) são evidenciados por uma unidade fundamental dos genomas bacterianos. Essa unidade é o *operon*, um intervalo de genes na sequência gênica de um dado genoma, que é recorrente nas sequências de outros genomas (ESCH; MERKL, 2020). Esses *operons* são decorrentes da pressão seletiva que as proteínas interagentes sofrem durante a evolução das espécies, a fim de se manterem próximas na fita de DNA. Geralmente, proteínas codificadas no mesmo *operon* trabalham juntas para desempenhar funções específicas ou estão envolvidas no mesmo processo biológico (ESCH; MERKL, 2020). Proteínas que formam um complexo proteico por meio de interações físicas ou trabalham juntas na mesma via metabólica, costumam ser codificadas no mesmo *operon* em diferentes genomas (MERING et al., 2003). A ordem dessas proteínas é conservada entre os diferentes genomas, embora a estrutura do *operon* seja instável durante a evolução das espécies (MERING et al., 2003). Baseando-se nessa evidência, a predição computacional de PPIs por vizinhança gênica conservada, infere interações entre proteínas a partir da contatação de *operons* nos genomas de interesse (Figura 10). Eventos de vizinhança gênica conservada são um indicador confiável para inferir computacionalmente associações funcionais entre proteínas (ESCH; MERKL, 2020).

Eventos evolutivos que servem como evidências biológicas de interações entre proteínas, também incluem a **Fusão Gênica** (do inglês, Gene Fusion (GF)). Uma fusão de genes se refere ao evento em que dois genes vizinhos em um dado genoma, se fundem

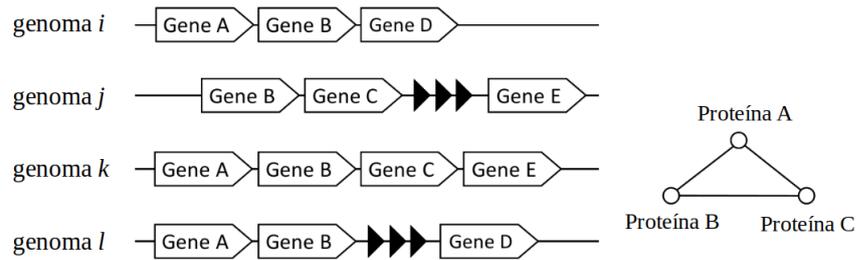


Figura 10 – Ilustração da predição de PPIs baseada em eventos de vizinhança gênica conservada. Infere-se que as proteínas A, B e C interagem entre si, pois elas ocorreram repetidamente próximas umas das outras (no mesmo *operon*) em pelo menos 2 do conjunto de 4 genomas analisados.

Fonte: Adaptado de (HIROSE, 2012)

como uma sequência contínua formando um novo gene em outro genoma (SNEL; BORK; HUYNEN, 2000). O gene fundido é chamado de Pedra de Rosetta (do inglês, Rosetta Stone). Esse método pressupõe que a fusão entre duas proteínas implique em uma interação física ou funcional de seus produtos proteicos (ENRIGHT et al., 1999). Por exemplo, a Figura 11 mostra que dois genes distintos (A e B) do genoma *i*, se fundiram formando o gene X no genoma *j*. Isso indica uma interação entre as proteínas A e B.

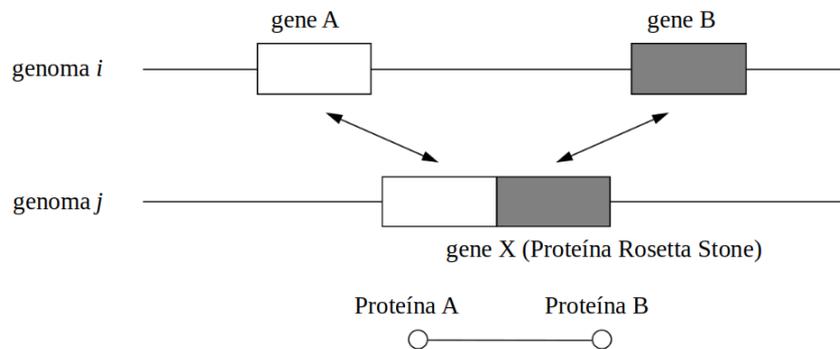


Figura 11 – Ilustração da predição de PPI baseada em eventos de fusão gênica. Infere-se que as proteínas A e B interagem entre si, pois foram fundidas em uma única proteína (proteína Rosetta Stone) em outro genoma.

Fonte: Adaptado de (HIROSE, 2012)

A fusão entre dois genes pode ocorrer como resultado de translocação, deleção intersticial ou inversão cromossômica (CHAREST et al., 2003). Computacionalmente, genes fundidos podem ser encontrados através de uma busca de similaridade de sequência de genes entre genomas. (ENRIGHT et al., 1999) aplicou esse método com sucesso a um grande número de genomas e foram previstos 39730 pares de proteínas associadas funcionalmente, oriundos de 24 genomas totalmente sequenciados. Uma grande vantagem desse

método é sua confiabilidade, porque eventos de fusão de genes são muito informativos sobre o relacionamento funcional. Uma desvantagem é que eventos de fusão gênica não são abundantes, principalmente em procariontes (PANCHENKO; PRZYTZYCKA, 2010).

Outro método de predição computacional de PPIs baseado nos eventos evolutivos das espécies, é o método de **Perfil Filogenético Conservado** (do inglês, Phylogenetic Profile (PP)). Assim como as abordagens de vizinhança gênica conservada e fusão gênica, esse método é sustentado pela hipótese geral de que as associações funcionais entre proteínas, podem ser inferidas com base nas relações evolutivas que elas possuem. A abordagem dos perfis filogenéticos proposta por (PELLEGRINI et al., 1999), sugere que durante o processo de evolução, pares de proteínas em interação serão preservados para funcionar em conjunto. Ou seja, se duas proteínas são necessárias para a realização de uma determinada função biológica, um organismo precisa possuir ambas as proteínas caso essa função seja necessária, enquanto ambas não fazem falta caso a função não seja necessária. Não faz sentido um organismo codificar apenas uma das duas proteínas em seu genoma. Baseando-se nessa hipótese, essa abordagem infere interações proteicas a partir dos perfis filogenéticos das proteínas. Um perfil filogenético de uma proteína representa simplesmente em quais organismos de um determinado conjunto de genomas de referência, existe um gene homólogo a essa proteína. Isto é, uma proteína com uma sequência de aminoácidos muito semelhante. O perfil filogenético de uma dada proteína é representado por uma cadeia de bits de tamanho igual ao número de genomas de referência. Cada bit corresponde a um genoma do conjunto. Se um dado bit é igual a 1, isso indica que a proteína possui um gene homólogo no genoma em questão, se é igual a 0 não possui um homólogo. (PELLEGRINI et al., 1999) mostram que, se duas proteínas de um dado genoma possuem genes similares nos mesmos genomas do conjunto analisado, ou seja, se duas proteínas possuem perfis filogenéticos idênticos, elas provavelmente interagem entre si. Eles também mostram evidências de apoio a essa hipótese, pois na medida em que as proteínas que se pensa não interagirem, estas não apresentam perfis filogenéticos semelhantes.

O primeiro passo para aplicação desse método, é verificar a existência de similaridade entre todas as proteínas de um genoma de consulta e as de outros genomas de referência. Para tanto, utiliza-se algoritmos para comparação par-a-par sequências de aminoácidos de proteínas, como o BLAST por exemplo. Simultaneamente, constrói-se um perfil filogenético para cada proteína do genoma de consulta. Esse perfil é constituído de valores binários como 1 para indicar a presença de um gene homólogo em algum genoma de referência, e 0 para indicar a ausência. Assim, a dimensão do perfil filogenético de uma proteína do genoma de consulta, é o número de genomas de referência. Finalmente, infere-se que os pares de proteínas com perfis idênticos ou semelhantes, interagem entre si (mais precisamente, são funcionalmente associados) (PELLEGRINI et al., 1999). A escolha dos genomas de referência é crucial para essa abordagem (SUN; LI; ZHAO, 2007). Além

disso, a exatidão do algoritmo utilizado para identificar similaridade de sequência entre as proteínas, afeta fortemente a montagem dos perfis e, conseqüentemente, a confiabilidade das previsões desse método. Uma ilustração dessa abordagem é dada pela Figura 12.

Genes de Consulta	Organismos de Referência				
	R ₁	R ₂	R ₃	R ₄	R ₅
gene A	1	0	1	0	1
gene B	1	1	0	1	1
gene C	1	0	1	0	1
gene D	1	0	1	1	1
gene E	0	1	1	0	1

Proteína A Proteína C



Perfil Filogenético (10101)

Figura 12 – Ilustração da predição de PPI baseada em eventos de perfil filogenético conservado. Infere-se que as proteínas A e C interagem entre si, pois possuem o mesmo perfil filogenético (10101). Isto é, elas ocorrem nos mesmos genomas e estão ausentes nos mesmos genomas de referência.

Fonte: Adaptado de (HIROSE, 2012)

2.2.5.3 Abordagens de Aprendizado de Máquina

Resumidamente, métodos baseados em aprendizado de máquina treinam um classificador binário usando interações proteicas conhecidas para discernir entre pares de proteínas interagentes e não interagentes de amostras de consulta (QI et al., 2010). O KNN (K vizinhos mais próximos), NB (Naive Bayesian), SVM (Máquina de Vetores de Suporte), DNN (Redes Neurais Profundas), e RF (Random Forest), são algoritmos de aprendizado de máquina muito utilizados nesse problema (DEY; MUKHOPADHYAY, 2019) (ZENG et al., 2020) (YANG et al., 2020). A vantagem das abordagens de aprendizado de máquina em relação ao método *ab initio* apresentado na subseção anterior, é o tempo de processamento significativamente menor. Por outro lado, o ponto fraco é que o desempenho preditivo varia amplamente dependendo da qualidade do conjunto de dados de treinamento e da seleção de métodos estatísticos apropriados. Para uma previsão de PPIs bem sucedida através de abordagens de aprendizado de máquina, o passo chave é a seleção de um conjunto de características adequado para representar proteínas. Por exemplo, a sequência de aminoácidos da proteína, informações sobre a estrutura da proteína, níveis de co-expressão gênica, perfis filogenéticos, informações evolutivas e afinidades de ligação conhecidas, são usadas como características para treinar um preditor de interações proteicas (ZHENG et al., 2019). As características adotadas devem ser convertidas em vetores numéricos de dimensão fixa para serem os dados de entrada do classificador.

Dois pontos devem ser considerados quando um modelo de previsão é construído. O primeiro diz respeito à necessidade de prestar atenção na qualidade dos dados utilizados para o treinamento, uma vez que o desempenho do modelo preditivo depende fortemente deles. O segundo enfatiza a importância de escolher o algoritmo de aprendizado de máquina mais apropriado para o problema em questão.

De acordo com (WANG et al., 2020), as abordagens de aprendizado de máquina para predição de interações entre proteínas, apresentam desafios técnicos diversos dentre os quais destaca-se:

- ❑ Como estabelecer um conjunto de dados abrangente, preciso e confiável para treinar um classificador?
- ❑ As proteínas têm uma grande variedade de propriedades físicas e químicas e diferentes características estruturais. Portanto, a extração eficiente e precisa de características para representar uma proteína, é um problema comum enfrentado na predição de PPIs por algoritmos de aprendizado de máquina.
- ❑ Como combinar as características de duas proteínas e converter em vetores numéricos para representar pares de proteínas?
- ❑ Como escolher um algoritmo de aprendizado de máquina eficiente e preciso, que faça pleno uso das características adotadas e gere um modelo preditivo eficaz?
- ❑ Alguns algoritmos de aprendizado de máquina são fáceis de apresentar sobre-ajuste (do inglês, *overfitting*). O resultado é um classificador muito bem ajustado para um conjunto de genomas específico, mas ineficaz para prever interações entre proteínas de outros genomas.
- ❑ A maioria dos modelos de previsão de PPIs baseados em aprendizado de máquina existentes, são treinados utilizando conjuntos de dados balanceados. Porém, conjuntos de dados de interações proteicas realistas geralmente são desbalanceados, o que leva ao treinamento de um preditor com "preferência".

Além desses desafios técnicos mencionados por (WANG et al., 2020), vale lembrar que em cada novo genoma elucidado, pelo menos 10% de seus genes serão exclusivos daquele genoma específico (LAPIERRE; GOGARTEN, 2009). Isso significa que não haverá um padrão de PPI previamente mapeado em qualquer modelo preditivo de aprendizado de máquina para esses genes inéditos. Assim, o preditor gerado provavelmente os classificaria como dados ruidosos negligenciando 10% de possíveis interações de proteínas. Outra questão, é que quase todos os modelos preditivos construídos são projetados para certos gêneros bacterianos específicos, limitando sua generalização a outros agentes infecciosos.

2.3 Trabalhos Correlatos

A predição *ab initio* de redes de PPI a partir de dados genômicos baseados nos eventos evolutivos das espécies, pode se tornar uma tarefa computacionalmente dispendiosa, tanto em termos de tempo de processamento quanto em memória.

Abordagens de aprendizado de máquina tem sido aplicadas frequentemente nos últimos anos à tarefa de predição de interações entre proteínas. Modelos preditivos com níveis de acurácia satisfatórios, tem sido gerados. Várias tecnologias avançadas de redes neurais foram introduzidas anteriormente e seu potencial para lidar com problemas que envolvem dados de sequências biológicas foi brevemente explorado. Recentemente, verificou-se a capacidade da tecnologia de aprendizado profundo no campo da bioinformática, e sua aplicação na predição de PPIs aumentou ano a ano. (YAO et al., 2019) utilizam uma rede neural profunda para esse problema de predição. Os autores propuseram um novo método de representação de resíduos de aminoácidos chamado Res2vec para a representação da sequência de proteínas. As representações de resíduos de aminoácidos obtidas pelo Res2vec descrevem com mais exatidão a sequência bruta das proteínas e fornecem entradas mais efetivas para o modelo de aprendizado profundo. O modelo preditivo proposto denominado DeepFE-PPI, é avaliado nos conjuntos de dados *S. Cerevisiae* e *humano*. Os resultados experimentais mostram que o DeepFE-PPI atingiu 94,78% de acurácia em suas predições. Os autores não disponibilizam o DeepFE-PPI como uma ferramenta de predição para acesso público, mas fornecem os códigos e instruções necessárias para reprodução desse trabalho na plataforma GitHub: <<https://github.com/xal2019/DeepFE-PPI>>.

Um trabalho que resultou na disponibilização de uma nova ferramenta de acesso público para predição de PPIs, é o trabalho realizado por (ROMERO-MOLINA et al., 2019). Nesse trabalho os autores apresentam um novo procedimento para a codificação numérica de polipeptídeos de propósito geral. Esse procedimento transforma pares de sequências de aminoácidos em um vetor de entrada amigável ao aprendizado de máquina. Os elementos do vetor são descritores numéricos de resíduos de proteínas. Os autores utilizaram esse procedimento de codificação numérica para o desenvolvimento de um classificador SVM que permite prever se duas proteínas interagem ou não. O modelo preditivo foi denominado PPI-Detect e está disponível para acesso público no site: <<https://ppi-detect.zmb.uni-due.de/>>. Para treinar e testar o PPI-Detect, os autores reuniram um conjunto de dados de benchmarking não redundante de PPIs com curadoria. Esse conjunto de dados é proveniente de três bancos de dados abrangentes, disponíveis publicamente. Esses bancos de dados contêm informações sobre pares de proteínas com interações comprovadas, e de pares não-interagentes. No total, os autores reuniram 4327 pares de proteínas, sendo 1922 pares interagentes e 2405 pares não-interagentes. Foram utilizados 3491 pares de proteínas (1613 interagentes e 1878 não-interagentes) na fase de treinamento, e 836 pares (309 interagentes e 527 não-interagentes) para testes. Os autores

comparam o desempenho do PPI-Detect com preditores do estado da arte, como o PIPE, SPPS, e Pred-PPI. PIPE é um preditor baseado em alinhamento de sequência, que realiza avaliações massivas da probabilidade de interação entre pares de proteínas. Para realizar suas predições, o PIPE considera todas as interações conhecidas de várias espécies, como por exemplo, *humano*, *levedura* e *Escherichia coli*. Já o Pred-PPI e SPPS são preditores de SVM. O modelo preditivo proposto pelos autores apresentou o melhor desempenho, tendo obtido 66% de acurácia, seguido pelo preditor PIPE com 63%, SPPS com 61% e o Pred-PPI com 43%.

Dentre as opções de bancos de dados públicos que fornecem informações sobre interações de proteínas, destaca-se a ferramenta de pesquisa para recuperação de genes/proteínas em interação (STRING) (SZKLARCZYK et al., 2019). O STRING apresenta dados de anotação para mais de cinco mil genomas, espalhados por uma gama significativa de organismos. Características como vizinhança gênica conservada, perfil filogenético conservado, fusão de genes, características de ontologia genética (função molecular, processo e localização), coexpressão gênica, experimentos bioquímicos e evidências bibliográficas são conjugadas para criar uma força probabilística de crença (score) de interações inferidas para pares de proteínas. Essa ferramenta realiza a predição de interações entre proteínas usando um método tradicional, a similaridade de sequência. Os usuários fazem consultas ao banco de dados do STRING por meio de uma ou mais proteínas de interesse, para as quais associações funcionais devem ser previstas. Essas proteínas podem ser identificadas por seus números de acesso ou identificador. Alternativamente, a sequência bruta de aminoácidos da proteína pode ser fornecida. Neste caso, são feitas buscas por similaridade de sequência entre as proteínas de consulta do usuário e as proteínas cadastradas no banco de dados do STRING. O objetivo é identificar genes homólogos presentes no banco de dados. Assim, se as proteínas de consulta do usuário não estiverem presentes no banco de dados do STRING, ou se no mesmo não existir proteínas homólogas às proteínas de consulta, não haverá uma predição de PPI para essas proteínas. Mas, se houver proteínas homólogas cadastradas no banco de dados do STRING, então o mesmo apresenta ao usuário um relatório das ligações funcionais previstas para as proteínas de consulta. Esse relatório conta com um score de confiança estimado para cada interação prevista. Outras informações que resumem e explicam as evidências que levam às previsões hipotéticas, também são fornecidas. Além disso, uma tela da rede de PPI predita, totalmente interativa, é disponibilizada permitindo a navegação pelas associações funcionais previstas. Para uma análise computacional independente, todo o conjunto de previsões de redes de interação contidas no banco de dados do STRING, é disponibilizado pelo site em arquivos simples de texto. STRING está disponível para acesso público em: <<https://string-db.org/>>.

Para um exemplo de estudo envolvendo a ferramenta STRING, podemos citar (SUN; ZHANG, 2020). Nesse trabalho os autores estudaram genes cruciais do câncer carcinoma hepatocelular (HCC). Os dados iniciais foram obtidos da base de dados Gene Expression

Omnibus (CLOUGH; BARRETT, 2016), compreendendo 93 amostras com tumor e 41 amostras saudáveis. O pacote *limma* do Rstudio foi aplicado para identificar genes diferencialmente expressos (DEGs) entre HCCs (amostras com tumor) e amostras saudáveis. Para explorar as funções biológicas dos HCC-DEGs (genes diferencialmente expressos de amostras com tumor), os autores utilizaram o software online DAVID realizando análises de enriquecimento via GO (Gene Ontology) (CONSORTIUM, 2019) e KEGG (Kyoto Encyclopedia of Genes and Genomes) (KANEHISA et al., 2019). DAVID é um site de análise para anotação, visualização e descoberta integrada de DEGs (JIAO et al., 2012). A ferramenta STRING foi usada para identificar prováveis interações de proteínas. Os autores carregaram os HCC-DEGs no banco de dados STRING e uma pontuação de interação mínima igual a 0,9 foi definida visando resgatar apenas interações proteicas com 90% de probabilidade de interação ou mais. Posteriormente, o software Cytoscape (LOPES et al., 2010) foi usado para visualizar e analisar as redes de interação preditas pelo STRING. Uma vez no Cytoscape, o plugin MCODE foi usado para filtrar módulos importantes de toda a rede de PPI, a fim de identificar genes importantes entre os HCC-DEGs. Para finalizar, eles usaram o site Gene Expression Profiling Interactive Analysis (GEPIA) (TANG et al., 2017) para determinar os efeitos dos genes identificados, na sobrevida geral do câncer carcinoma hepatocelular. O trabalho de (SUN; ZHANG, 2020) é uma conjugação bastante elaborada de vários bancos de dados e ferramentas de bioinformática para produzir análises *in silico* (computacionais) interessantes. Existem muitos outros estudos semelhantes aos citados nesta seção. Uma simples busca no google acadêmico para os principais termos aqui mencionados, poderia recuperar uma miríade de trabalhos relacionados.

Proposta

Este capítulo apresenta a proposta deste trabalho e discorre sobre a implementação da mesma. Vale mencionar que aqui, o termo **gene** tem o mesmo significado semântico do termo **proteína**.

3.1 Introdução

Conforme enfatizado na Subseção 1.1.2 referente ao estado da arte da predição computacional de redes de interação entre proteínas, ainda há problemas não resolvidos nessa área de pesquisa. Tanto as alternativas de bancos de dados públicos que fornecem informações sobre interações de proteínas conhecidas (STRING, por exemplo), quanto as abordagens de aprendizado de máquina, apresentam uma mesma limitação, a impossibilidade de predição de PPIs para genes inéditos. Essa limitação é decorrente do fato dessas alternativas utilizarem proteínas e interações conhecidas para embasar suas predições. Consequentemente, ocorre o negligenciamento das proteínas inéditas de novos genomas sequenciados e montados. Proteínas essas que representam pelo menos 10% dos genes desses genomas (LAPIERRE; GOGARTEN, 2009). Tal limitação pode implicar em milhares de possíveis interações negligenciadas.

Uma maneira de resolver esse problema é através do método *ab initio* para predição de PPIs baseada em dados genômicos. Esse método não utiliza conhecimento *a priori* de outras interações proteicas conhecidas, mas realiza suas predições a partir de evidências biológicas que podem ser encontradas nos próprios genomas. Essas evidências são rastros deixados por eventos evolutivos que indicam interações entre proteínas. Ao analisar computacionalmente vários genomas em conjunto, podemos identificar esses rastros e inferir interações confiáveis de proteínas. Para tanto, faz-se necessário uma ferramenta eficaz tanto em termos de tempo de processamento quanto em exatidão, para comparar para-a-par, sequências de aminoácidos de proteínas. No entanto, utilizando as alternativas atuais como o algoritmo heurístico da ferramenta BLAST ou o algoritmo exato Needleman-Wunsch, essa solução se torna impraticável. O motivo é que tais algoritmos

são computacionalmente caros e tornam esse método de predição *ab initio* muito dispendioso tanto em termos de tempo de processamento quanto em consumo de memória. Acreditamos que esse seja o motivo da indisponibilidade atual de soluções computacionais para prever interações de proteínas através de um processo algorítmico *ab initio*.

Para contornar o obstáculo referente ao tempo de processamento, propomos o GenPPI, um novo software autônomo de bioinformática capaz de processar um número expressivo de genomas em um tempo aceitável, mesmo que utilizando uma máquina de configuração convencional. A título de exemplo, para 50 genomas contendo em média 2.200 proteínas, essa ferramenta realiza suas predições com menos de uma hora de processamento. Esse processamento rápido é possível graças a uma nova heurística para comparação par-a-par de sequências de aminoácidos de proteínas, introduzida por este trabalho. O software proposto inspeciona genomas representados por arquivos multi-fasta de proteínas, fazendo inferências de PPIs a partir de evidências de eventos evolutivos que indicam interações de proteínas. Os eventos averiguados são os de vizinhança gênica conservada, fusão gênica e perfil filogenético conservado. O GenPPI é um software autônomo disponibilizado na forma de um arquivo executável para ser utilizado através de parâmetros informados na linha de comando de um terminal Linux ou do prompt de comando do Windows.

Com essa proposta contribui-se disponibilizando para acesso público, uma nova ferramenta de bioinformática para predição *ab initio* de redes de interação entre proteínas bacterianas, a partir dos próprios genomas de interesse do usuário. Assim, é suprida uma deficiência identificada no estado da arte, a indisponibilidade de soluções computacionais capazes de realizar predições de interações proteicas, sem negligenciar as proteínas inéditas dos novos genomas. Além disso, contribui-se também introduzindo uma nova heurística para comparação par-a-par de sequências de aminoácidos de proteínas. Com essa heurística, conseguimos fazer comparações de pares de proteínas, em um tempo significativamente menor que o principal algoritmo heurístico da atualidade (BLASTp). Sem deixar a desejar em termos de exatidão na comparação par-a-par de proteínas. Isso será demonstrado na Subseção de resultados 4.2.2. Acredita-se que tal heurística é uma alternativa viável frente a utilização de um algoritmo exato ou até mesmo das melhores soluções heurísticas disponíveis atualmente.

O restante deste capítulo segue assim: a Seção 3.2 apresenta as tecnologias utilizadas na implementação da proposta; a Seção 3.3 apresenta o fluxograma do software desenvolvido; a Seção 3.4 fala sobre os dados de entrada do programa que são arquivos multi-fasta de proteínas oriundas de genomas bacterianos; a Seção 3.5 demonstra o funcionamento de nossa heurística para comparação par-a-par de proteínas, e descreve a implementação da primeira etapa de uma execução do GenPPI. Essa etapa é a geração de histogramas de aminoácidos para aplicação da heurística do programa; a Seção 3.6 descreve a implementação da segunda etapa referente à geração do Pan-Genoma dos organismos envolvidos em uma análise; a Seção 3.7 descreve a implementação da terceira etapa que é uma predição

ab initio de interações entre proteínas, com base na constatação de eventos evolutivos de vizinhança gênica conservada, fusão gênica e perfil filogenético conservado; a Seção 3.8 descreve a fase de encerramento de uma execução do GenPPI; a Seção 3.9 apresenta a complexidade algorítmica do programa; e a Seção 3.10 descreve uma otimização de código feita para resolver problemas de estouro de memória.

3.2 Tecnologias Utilizadas no Desenvolvimento da Proposta

LISP (“List Processing”) é uma linguagem de programação criada nos fins da década de 50 pelo matemático John McCarthy, e que introduziu o conceito de processamento simbólico. De maneira análoga ao Prolog, o Lisp é uma linguagem de programação adequada para programas de inteligência artificial, principalmente porque seus recursos exclusivos permitem o processamento eficaz de informações simbólicas. Na ferramenta computacional desenvolvida neste trabalho, não há muito de processamento de símbolos, mas foi implementada uma heurística para comparações de sequências de aminoácidos de proteínas. Em conformidade com a Subseção 2.2.2 do capítulo de fundamentação teórica, o Lisp se destaca na IA clássica. Porém, comparado a Python que é muito utilizado para aprendizado de máquina, o Lisp não pode lidar com demandas desse tipo. A razão para isso é a falta de suporte na forma de bibliotecas. Independente disso, o treinamento em Lisp faz um programador melhor para lidar com problemas complexos, pois oferece recursos de programação inexistentes em outras linguagens.

O principal diferencial do Lisp é o seu sistema exclusivo de macros que permite estender a sintaxe dentro do próprio Lisp. Ou seja, introduzir novas construções de linguagem e sintaxe descrevendo-as no próprio Lisp. Isso significa que uma nova construção de sintaxe pode ser adicionada semelhantemente a uma biblioteca, sem precisar alterar o compilador. Além das macros, o Lisp também possui uma grande variedade de funções na sua biblioteca padrão que tornam essa linguagem de programação mais flexível do que muitas outras e facilitam o desenvolvimento de programas complexos.

Considerando as vantagens do Lisp, o mesmo foi escolhido para desenvolver a proposta deste trabalho. O GenPPI foi desenvolvido em ambiente Linux programando no paradigma de programação procedimental que especifica os passos que um programa deve seguir para alcançar um objetivo desejado. Como editor de código-fonte utilizou-se o Atom em sua versão 1.50.0. Para controle de versões foi utilizado a plataforma de hospedagem de código-fonte Bitbucket. A seguir é apresentado o fluxograma da solução proposta.

3.3 Fluxograma do GenPPI

As predições computacionais feitas pela solução proposta são divididas nas 3 etapas principais que estão demonstradas no fluxograma da Figura 13.

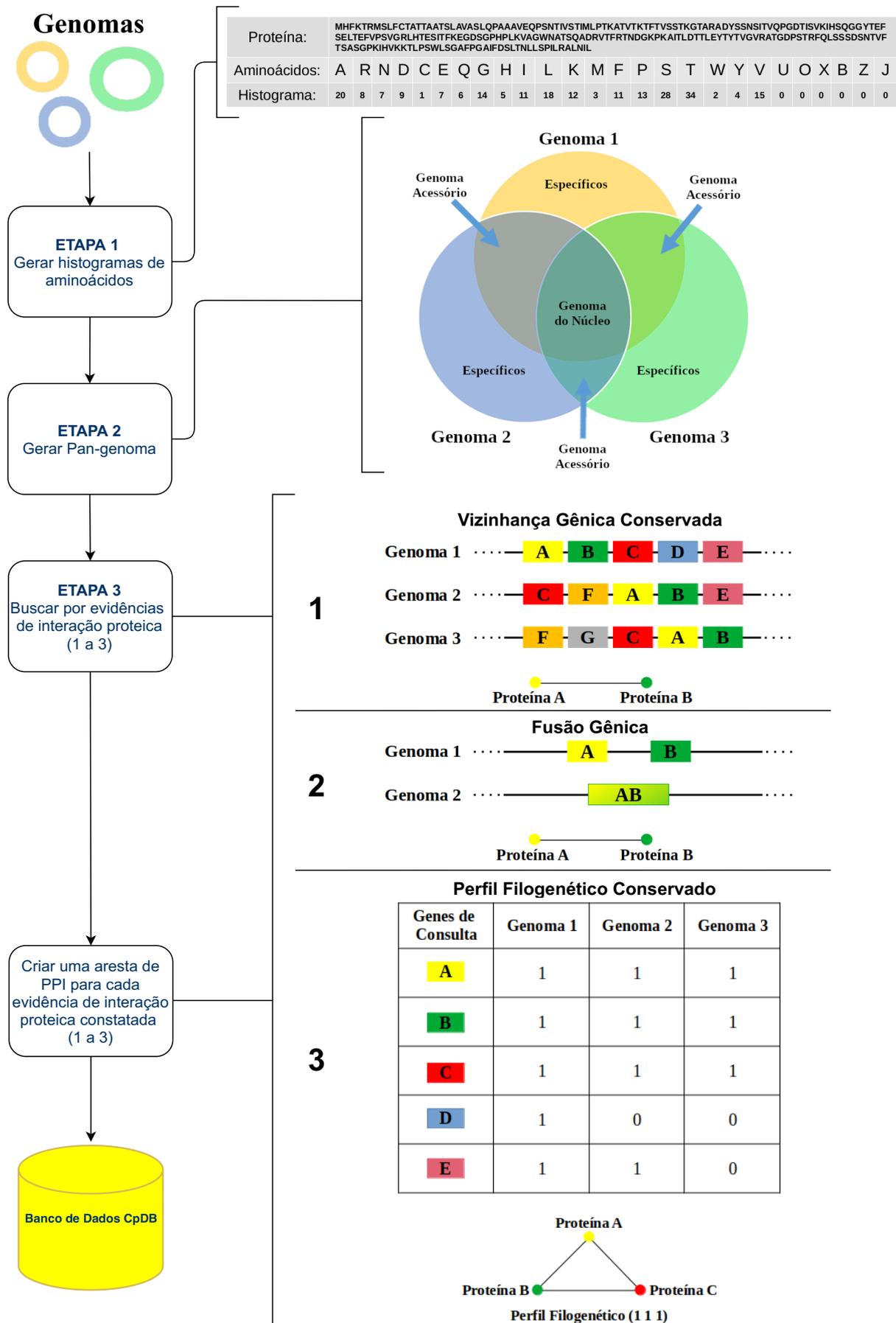


Figura 13 – Fluxograma do GenPPI

3.4 Dados de Entrada do GenPPI

O GenPPI não trabalha com a fita de DNA, mas com um relatório exportado a partir do DNA, que são as sequências de aminoácidos das proteínas. O programa considera que as proteínas recebidas via um arquivo multi-fasta, estão ordenadas tal qual os seus genes estão dispostos em uma fita de DNA. Parti-se da premissa que as proteínas entram em um arquivo multi-fasta em uma sequência similar a que estavam quando foram extraídas da fita de DNA.

Um arquivo multi-fasta consiste em nomes de proteínas e sequências de aminoácidos. Os identificadores são precedidos por um símbolo de maior que (>) e o restante da linha é uma descrição da proteína em questão. As próximas linhas até o próximo símbolo de maior que, contêm a sequência real de aminoácidos de uma proteína. A Figura 14 mostra um exemplo de proteína em um arquivo multi-fasta. O símbolo de maior que (>) é seguido pelo nome da proteína (YPR161C). A sequência real de aminoácidos da proteína é dada nas próximas cinco linhas.

```
> Y P R 1 6 1 C
T T G A C M T T G A C M T T G A C M V V V R N M
A C M T T G A C T T G A C M T T G M T T G A C M
Q S D A V M T T G A C M T T G A C M T T G A C M
A C M T T G A C M T T G A C M T T G T T G A C M
T T G A C M T T G A C M T T
```

Figura 14 – Uma proteína em um arquivo multi-fasta

Um arquivo multi-fasta não possui apenas uma proteína, mas todas as proteínas de um organismo, que podem totalizar centenas ou até milhares de proteínas em um organismo bacteriano. Esses arquivos multi-fasta com as sequências de aminoácidos de proteínas oriundas de bactérias, podem ser obtidos através do banco de dados público do NCBI.

3.5 Etapa 1 - Gerar Histogramas de Aminoácidos Para Aplicação da Heurística do GenPPI

Para que seja possível o emprego dos métodos de inferência de PPI baseados em evidências biológicas no contexto genômico, faz-se necessário um algoritmo eficaz para classificar se um par de proteínas é similar ou não. No entanto, o algoritmo exato para comparação de sequências biológicas Needleman-Wunsh, possui complexidade computacional quadrática tanto em termos de tempo quanto em espaço ($\mathcal{O}(mn)$ onde m e n representam os tamanho das sequências a serem comparadas). Essa complexidade torna impraticável o emprego do método de predição *ab initio* de redes de PPI a partir de dados genômicos.

Mesmo utilizando as alternativas de algoritmos heurísticos para comparação de sequências de proteínas, como o BLAST por exemplo, esse método se torna impraticável em decorrência de um longo tempo de processamento.

Para contornar esse problema, neste trabalho propõe-se uma heurística baseada no histograma de aminoácidos das proteínas. Com o emprego dessa heurística é possível fazer comparações par-a-par de proteínas, com uma complexidade computacional linear ($\mathcal{O}(n)$). Vale mencionar que n sempre será igual a 26, o número de possíveis tipos diferentes de aminoácidos em uma proteína. Assim, para comparar duas proteínas, a nossa heurística realiza apenas 26 comparações referentes às frequências dos 26 possíveis tipos de aminoácidos nas sequências do par de proteínas comparado. Consequentemente, essa heurística realiza comparações par-a-par de sequências em um tempo significativamente menor quando comparado ao tempo gasto pelos algoritmos clássicos de alinhamento de sequências biológicas. Além disso, não é inferior em termos de exatidão na comparação par-a-par de proteínas.

O primeiro processo realizado em uma execução do programa desenvolvido, é a geração de histogramas de aminoácidos para todas as proteínas dos genomas incluídos em uma análise. Para tanto, todos os arquivos multi-fasta de proteínas dos genomas que serão analisados, devem ser inseridos em um diretório. Depois, no momento da execução do programa, esse diretório deve ser informado. Assim esses arquivos são lidos de modo a calcular a distribuição de frequência (histograma) dos 26 possíveis tipos de aminoácidos para todas as proteínas dos arquivos multi-fasta. Os histogramas de aminoácidos gerados são armazenados em estruturas de dados do tipo lista. Desse modo, uma proteína é representada por uma lista de 26 valores que são seu histograma de aminoácidos. A Tabela 3 demonstra duas proteínas A e B representadas nesse formato. Para armazenar todas as listas de histogramas de aminoácidos referentes às proteínas de um dado genoma, foi utilizado outra estrutura de dados do tipo lista. Essa nova lista é incluída em uma tabela-hash denominada *genomas* utilizando o nome do genoma em questão, para ser a chave de acesso a esse registro. No final desse processo, conta-se com a tabela-hash *genomas* contendo em seus registros, as listas de histogramas de aminoácidos das proteínas oriundas de cada genoma de um conjunto analisado. Um exemplo ilustrativo dessa tabela-hash pode ser visto na Figura 15.

Posteriormente, para comparar duas proteínas aplica-se a heurística do GenPPI. Para tanto, inicialmente é calculada a diferença dos histogramas de aminoácidos do par de proteínas a ser comparado. Depois, para classificar essas proteínas como sendo similares ou não, a heurística se baseia na diferença dos histogramas de aminoácidos desse par de proteínas, considerando dois parâmetros: o limite tolerado na diferença de histogramas (*d-limite*) e a quantidade de aminoácidos dentro desse limite de diferença tolerado (*qtd-amin*). A título de exemplo, a Tabela 3 apresenta uma amostra de duas proteínas (A e B) para comparação. Para classificar se esse par é similar ou não, a heurística do programa

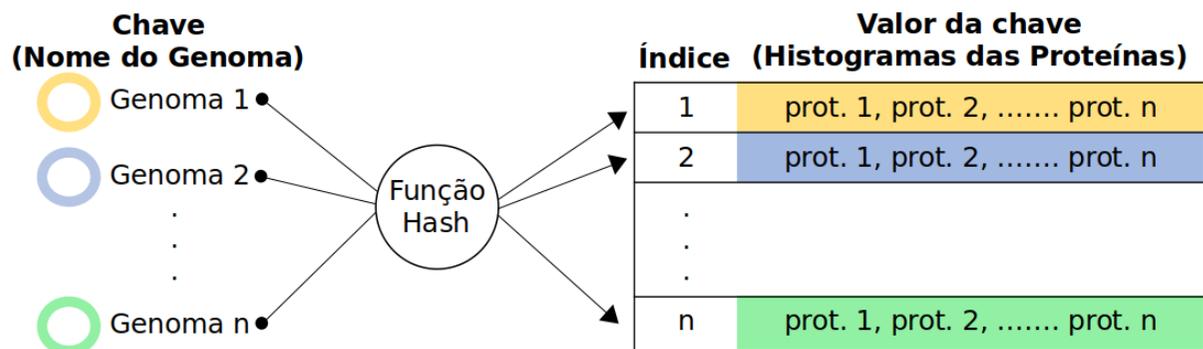


Figura 15 – Ilustração da tabela hash *genomas*

Tabela 3 – Heurística para comparação par-a-par de proteínas.

Amostra de duas proteínas para comparação																										
Proteína A	MAYSKKVMMDHYENPRNVGFSFSNSDNNVGSGLVGAPACGDVMKLQIKVNEKGIIEDACFKTYGCGSAI ASSSLVTEWVKGKSITEAESIRNTTIVEELELPPVKIHCSILAEDAIAAIAADYKSKKYSN																									
Proteína B	MAYSKKVMMDHYENPRNVGFSFSNSDLNVGSLVGAPACGDVMKLQIKVNEEGIIEDACFKTYGCGSAI ASSSLVTEWVKGKSIVEAESIRNTTIVEELELPPVKIHCSILAEDAIAAISDYKRKKNLN																									
Histograma de aminoácidos																										
Aminoácidos	A	R	N	D	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V	U	O	X	B	Z	J
Histograma A	12	2	8	6	4	10	1	9	2	11	6	13	3	2	4	14	5	1	5	10	0	0	0	0	0	0
Histograma B	11	3	8	6	4	11	1	9	2	11	8	12	3	2	4	13	4	1	4	11	0	0	0	0	0	0
Diferença AB	1	1	0	0	0	1	0	0	0	0	2	1	0	0	0	1	1	0	1	1	0	0	0	0	0	0
Classificação:	Segundo a heurística proposta as proteínas A e B são similares , pois na diferença de seus histogramas de aminoácidos, existe pelo menos 25 dentre os 26 possíveis tipos, cujas diferenças de valores de histograma é de no máximo 1. Com os valores 1 e 25 para os parâmetros <i>d-limite</i> e <i>qtd-amin</i> , respectivamente, essa heurística garante um percentual de identidade mínimo igual a 92.55% para os pares de proteínas classificadas similares. Essa garantia é assegurada pelos resultados apresentados na Tabela 8.																									

calcula a diferença de seus histogramas de aminoácidos. Em seguida, é verificada a quantidade de aminoácidos com diferença de frequência dentro do limite tolerado. Caso houver no mínimo 25 aminoácidos (parâmetro *qtd-amin*) dentre os 26 possíveis tipos, cujas diferenças de frequência seja de no máximo 1 (parâmetro *d-limite*), então as proteínas A e B são classificadas similares pela heurística.

3.6 Etapa 2 - Gerar Pan-genoma

Após a geração de histogramas de aminoácidos para as proteínas dos genomas incluídos em uma análise, a próxima etapa é gerar o pan-genoma desses organismos. De acordo com o que foi explicado na Subseção 2.1.7 do capítulo de fundamentação teórica, um pan-genoma é todo o conjunto de genes para todas as linhagens de um clado (grupo de organismos que descendem de um ancestral comum). Conforme ilustrado pela Figura 4, um pan-genoma inclui: o genoma do núcleo contendo genes presentes em todas as linhagens do clado, o genoma acessório contendo genes presentes em um subconjunto das linhagens, e genes específicos de cada linhagem.

No GenPPI, um pan-genoma é um mapeamento que registra a localização de todas as proteínas conservadas (proteínas presentes em duas ou mais linhagens do conjunto total de genomas analisado). Esse mapeamento é feito verificando a similaridade de uma dada proteína em relação a todas as outras através da heurística proposta. Nessa verificação, registra-se a localização exata das proteínas similares à proteína comparada. A localização é o nome do genoma e a posição nesse genoma onde foi identificada uma proteína similar. Como já mencionado anteriormente, um genoma é representado por uma estrutura de dados do tipo lista na qual um elemento é uma proteína. Cada proteína, por sua vez, é representada por outra lista contendo o histograma de aminoácidos dessa proteína. Assim, a localização exata de uma proteína consiste em dois dados: o nome do genoma (chave de acesso à lista de proteínas desse genoma na tabela-hash *genomas*), e o índice dessa lista no qual a proteína em questão está armazenada, por exemplo: (genoma-x, 577).

Após a geração do pan-genoma, é possível saber em quais linhagens do conjunto analisado, uma determinada proteína conservada ocorre e a localização exata (posição) dessa proteína na sequência gênica das linhagens nas quais ela está presente. Tal mapeamento é de suma importância, pois para verificar a existência de eventos de vizinhança gênica conservada e perfil filogenético conservado, é necessário saber a localização exata das proteínas conservadas, nas sequências gênicas dos genomas onde elas ocorrem. O pseudo-código da geração de um pan-genoma pelo programa desenvolvido, está demonstrado no Algoritmo 1.

A entrada do Algoritmo 1 é a tabela-hash *genomas* criada na etapa 1 (geração de histogramas de aminoácidos), e também os parâmetros da heurística. A saída é outra tabela-hash denominada *pangenoma* na qual os registros são mapeamentos que indicam a

Algoritmo 1 Gerar Pan-Genoma**Entrada:** tabela-hash *genomas*, *d-limite*, *qtd-amin*.**Saída:** tabela-hash *pangenoma*.

```

1: tabela-hash pangenoma ← vazio
2: lista auxiliar ← vazio
3: for all g ∈ genomas do
4:   for all proteína1 ∈ g do
5:     inserir-proteína-no-pangenoma(proteína1, pangenoma)
6:     for all proteína2 ∈ auxiliar do
7:       if heurística(proteína1, proteína2, d-limite, qtd-amin) then
8:         pangenoma.proteína1.similares ← proteína2.local
9:         pangenoma.proteína2.similares ← proteína1.local
10:      end if
11:    end for
12:    inserir-proteína-na-lista-auxiliar(proteína1, auxiliar)
13:  end for
14: end for
15: remover-proteínas-sem-genes-similares(pangenoma)
16: return tabela-hash pangenoma

```

localização das proteínas conservadas.

Inicialmente, nas linhas 1 e 2 a tabela-hash *pangenoma* e a lista *auxiliar* são inicializadas vazias. A lista *auxiliar* é utilizada na combinação de todos os possíveis pares de proteínas. Na sequência, os laços das linhas 3, 4 e 6 são responsáveis por gerar os $n^*(n-1)/2$ possíveis pares de proteínas. O valor n é o número total de proteínas contidas na tabela-hash *genomas* (todas as proteínas dos genomas incluídos em uma análise). Essas combinações de todos os possíveis pares de proteínas contidas nos genomas analisados, são necessárias para verificar entre quais proteínas existe similaridade de sequência de aminoácidos. O intuito é mapear a localização exata das proteínas que estão presentes (conservadas) em duas ou mais linhagens do conjunto de genomas analisado.

Na linha 5, antes de comparar uma proteína com todas as outras, cria-se um registro na tabela-hash *pangenoma* para armazenar a proteína que será comparada (utiliza-se o nome da proteína como chave de acesso). Esse registro contém dois campos: localidade e similares. O campo localidade é usado para armazenar o nome do genoma ao qual a proteína em questão pertence e o índice desse genoma onde ela está inserida. O campo similares é utilizado para armazenar a localização de eventuais proteínas similares à proteína comparada.

Para verificar se há similaridade entre um determinado par de proteínas, na linha 7 aplica-se a heurística explicada na seção anterior e demonstrada na Tabela 3. Os parâmetros dessa função heurística são: os histogramas de aminoácidos do par de proteínas a ser comparado, e os parâmetros da heurística (*d-limite* e *qtd-amin*). Condizente com a configuração padrão do programa, os parâmetros *d-limite* e *qtd-amin* estão definidos, res-

pectivamente, com os valores 1 e 25. Esses valores garantem um percentual de identidade mínima igual a 92,55% para os pares de proteínas classificadas similares. Tal garantia é assegurada pelos resultados obtidos na avaliação da heurística proposta (Subseção 4.2.2, Tabela 8).

O Algoritmo 2 apresenta o pseudocódigo da heurística proposta. Como dito anteriormente, os dados de entrada desse algoritmo são os histogramas de aminoácidos do par de proteínas a ser comparado, e os parâmetros *d-limite* e *qtd-amin*. A saída é booleana, ou seja, verdadeiro ou falso representando, respectivamente, se um determinado par de proteínas é similar ou não. O laço de 26 iterações da linha 2 calcula a diferença entre os histogramas de aminoácidos das duas proteínas, e conta quantos aminoácidos estão dentro do limite de diferença tolerado. Depois, na linha 7 do Algoritmo 2 é verificada a seguinte condição: se a quantidade de aminoácidos com diferença de valores de histograma dentro do limite tolerado (parâmetro *d-limite* = 1), atinge a quantidade mínima requerida pelo parâmetro *qtd-amin* (25). Finalmente, em concordância com a demonstração feita na Tabela 3, se tal condição for satisfeita, a heurística então classifica as duas proteínas em questão como sendo similares retornando o valor booleano verdadeiro. Do contrário, retorna falso indicando que esse par de proteínas não é similar.

Algoritmo 2 Heurística Histo-Fasta

Entrada: *proteína1*, *proteína2*, *d-limite*, *qtd-amin*.

Saída: verdadeiro ou falso.

```

1: checkpoint ← 0
2: for i = 0 to 25 do
3:   if diferença(proteína1[i], proteína2[i]) ≤ d-limite then
4:     checkpoint ← +1
5:   end if
6: end for
7: if checkpoint ≥ qtd-amin then
8:   retorna verdadeiro
9: else
10:  retorna falso
11: end if

```

Voltando ao Algoritmo 1 em sua linha 7, para exemplificar o que é feito após a função heurística classificar que um par de proteínas é similar, considera-se o exemplo das proteínas denominadas BAKON69 e BUAMB57 da Figura 16. Essas proteínas são oriundas de duas entre cinco cepas da bactéria *Buchnera Aphidicola*, utilizadas como exemplo.

Após a função heurística determinar que o par de proteínas BAKON69 e BUAMB57 é similar, na linha 8 do Algoritmo 1 é feito o que está descrito a seguir. Primeiramente, a proteína BAKON69 rebebe em seu registro da tabela-hash *pangenoma*, no campo similares, a localização da sua proteína similar BUAMB57 (Ba_Ua, 156). Essa localização é composta pelo nome do genoma ao qual a proteína BUAMB57 pertence e pelo índice

Chave de Acesso	Valor da Chave
BAKON69 =>	Localidade: (Ba_Ak 168) Similares: ((Ba_Ua 156))
BUAMB57 =>	Localidade: (Ba_Ua 156) Similares: ((Ba_Ak 168))

Figura 16 – Exemplo de dois registros da tabela-hash *pangenoma*, referentes a um par de proteínas classificadas como sendo similares pela heurística

desse genoma onde ela está inserida. Depois, na linha 9 o mesmo é feito para a proteína BUAMB57 tal qual ilustra a Figura 16.

Finalmente, após aplicar a função heurística a todos os possíveis pares de proteínas dos genomas incluídos em uma análise, conta-se com a tabela-hash *pangenoma* contendo genes do genoma do núcleo (proteínas presentes em todas as cepas analisadas, se houverem), genes do genoma acessório (proteínas presentes em um dado subconjunto de cepas do conjunto total) e genes específicos (proteínas presentes em apenas uma das cepas). De acordo com explicações anteriores, um registro da tabela-hash *pangenoma* (ex: Figura 17) contém os campos: localidade e similares. Estes representam, respectivamente, a localidade de uma proteína (nome e índice do genoma ao qual esta pertence) e uma lista da localidade de proteínas similares. Há proteínas (genes específicos) para as quais não se acham outras similares. Essas não são úteis na verificação de evidências biológicas de interações proteicas por vizinhança gênica conservada e perfil filogenético conservado, motivo pelo qual é necessário gerar um pan-genoma no GenPPI. Portanto, antes do Algoritmo 1 ser finalizado, na linha 15 tais proteínas são removidas da tabela-hash *pangenoma*.

3.7 Etapa 3 - Busca de Evidências Biológicas de Interações Proteicas no Contexto Genômico

Após a geração do pan-genoma dos organismos envolvidos em uma análise, a próxima etapa é a busca por evidências biológicas de interações entre proteínas nas sequências gênicas desses organismos. Para tanto, o programa desenvolvido conta com três abordagens de inferência de PPI baseadas em eventos evolutivos. Tais eventos englobam: vizinhança gênica conservada, fusão gênica e perfil filogenético conservado. Na sequência é apresentada a implementação dessas três abordagens.

3.7.1 Implementação da Predição de PPI Pelo Método de Vizinhança Gênica Conservada

Esse método é baseado na observação de que genes que ocorrem repetidamente próximos uns dos outros em genomas (*operons* – intervalos recorrentes de genes) tendem

a codificar proteínas que interagem funcionalmente (SNEL et al., 2000). Identificamos esses intervalos de genes verificando a vizinhança gênica de proteínas conservadas. Essa verificação é feita utilizando janelas de expansão que verificam a conservação gênica da vizinhança dos genes conservados. No algoritmo do GenPPI, os genes chamados conservados ou recorrentes são as proteínas da tabela-hash *pangenoma*. Portanto, a verificação da vizinhança gênica desses genes, é feita utilizando os registros dessa tabela-hash. Para fins de demonstração, a Figura 17 apresenta um registro da tabela-hash *pangenoma*. Esse registro é um mapeamento que indica a localização exata de uma proteína conservada (BAKON_493) em 4 genomas de um conjunto de 5 cepas da bactéria *Buchnera Aphidicola*, utilizadas como exemplo nesta subseção.

Chave de Acesso	Valor da Chave
BAKON_493 =>	Localidade: (Ba_Ak 491) Similares: ((Ba_G002 493) (Ba_Sg 497) (Ba_Ua 462))

Figura 17 – Exemplo de um registro da tabela-hash *pangenoma*

Cada registro da tabela-hash *pangenoma* é uma estrutura contendo dois campos: localidade e similares. No caso do exemplo da Figura 17, o campo localidade armazena a localização exata da proteína conservada BAKON_493, isto é, o nome do genoma ao qual essa proteína pertence e o seu índice no genoma. Semelhantemente, o campo similares armazena uma lista da localização das proteínas que possuem uma sequência de aminoácidos similar à sequência da proteína BAKON_493. Assim, o exemplo da Figura 17 indica que a proteína BAKON_493 localiza-se no índice 491 do genoma Ba_Ak, e que a mesma possui uma sequência de aminoácidos similar às sequências das proteínas localizadas nos índices 493, 497 e 462 dos genomas Ba_G002, Ba_Sg e Ba_Ua, respectivamente. Isso demonstra que a proteína BAKON_493 do genoma Ba_Ak, é uma proteína conservada e, portanto, passível de possuir genes vizinhos conservados junto com ela em outros genomas. Tal possibilidade caracteriza interações entre essas proteínas conservadas segundo o método de inferência de PPIs baseado em eventos evolutivos de vizinhança gênica conservada.

A identificação de genes conservados na vizinhança de uma proteína conservada, é feita comparando a vizinhança gênica dessa proteína com a vizinhança de suas proteínas similares de outros genomas. Desse modo, intervalos recorrentes de genes são identificados e inferências de PPIs são feitas entre esses genes. Para tanto, foram implementados dois métodos: expansão fixa e expansão dinâmica.

3.7.1.1 Expansão Fixa

O objetivo da expansão fixa é identificar intervalos recorrentes de genes tornando ajustável, por parte do usuário, a quantidade de genes a serem analisados na vizinhança de proteínas conservadas. A expansão fixa analisa a vizinhança gênica das proteínas conser-

vadas utilizando uma janela de tamanho fixo. As inferências de interações proteicas são feitas verificando a conservação dos genes da janela de expansão. O nível de conservação requerido para que inferências de PPIs sejam feitas, é dividido em 4 possíveis configurações. Essas configurações indicam intervalos decrescentes da janela de expansão e o nível de conservação gênica requerido dentro desses intervalos. Caso o usuário não informe essas configurações no momento de execução do programa, as mesmas assumirão seus valores padrões. O Algoritmo 3 demonstra o funcionamento da expansão fixa na inferência de PPIs baseada na métrica de vizinhança gênica conservada.

Algoritmo 3 Expansão Fixa

Entrada: *pangenoma*, *percentual-vgc*, *w1*, *cw1*, *w2*, *cw2*, *w3*, *cw3*, *w4*, *cw4*, *d-limite*, *qtd-amin*.

Saída: *tabela-hash ppi*.

```

1: for all proteína ∈ pangenoma do
2:   pivô-1 ← proteína.localidade
3:   lista conservação-gênica ← (1 1 1 1 1 1 1 1 1)
4:   for all pivô-2 ∈ proteína.similares do
5:     for i = 1 to w1 do
6:       if heurística(pivô-1+i, pivô-2+i, d-limite, qtd-amin) then
7:         incrementa(conservação-gênica[i-1])
8:       end if
9:     end for
10:  end for
11:  if conservação-gênica ≥ cw1 then
12:    criar-arestas-ppi(w1, percentual-vgc, ppi)
13:  else if conservação-gênica ≥ cw2 then
14:    criar-arestas-ppi(w2, percentual-vgc, ppi)
15:  else if conservação-gênica ≥ cw3 then
16:    criar-arestas-ppi(w3, percentual-vgc, ppi)
17:  else if conservação-gênica ≥ cw4 then
18:    criar-arestas-ppi(w4, percentual-vgc, ppi)
19:  end if
20: end for
21: normalizar-pesos(ppi)
22: return tabela-hash ppi

```

O algoritmo de expansão fixa varre toda *tabela-hash pangenoma* no laço *for all* da linha 1 comparando a vizinhança gênica das proteínas conservadas com a vizinhança de suas proteínas similares. Essa comparação é feita visando verificar a existência de níveis requeridos de conservação gênica na vizinhança das proteínas conservadas. Isso possibilita identificar intervalos recorrentes de genes nos genomas incluídos em uma análise. Sempre que um intervalo recorrente é constatado, são feitas inferências de PPIs para os pares de genes desse intervalo. Para exemplificar o funcionamento do Algoritmo 3, considera-se uma iteração do laço *for all* da linha 1, aplicada à proteína conservada BAKON_493

da Figura 17. Nesse exemplo, usa-se as configurações padrões para os parâmetros da expansão fixa. Esses parâmetros são referentes ao tamanho da janela de expansão e aos níveis de conservação gênica requeridos para se fazer inferências de interações de proteínas.

Inicialmente, é averiguado se a proteína conservada BAKON_493 faz parte de um intervalo recorrente de genes. Para tanto, entre as linhas 1 e 10 do algoritmo de expansão fixa, é feita a comparação da vizinhança gênica da proteína BAKON_493 (variável *pivô-1*) com a vizinhança de suas proteínas similares de outros genomas (variável *pivô-2*). Nessa comparação conta-se em quantos genomas cada um dos 10 genes vizinhos subsequentes da proteína BAKON_493, se mantiveram conservados na vizinhança gênica das proteínas similares a ela. Esse processo é ilustrado pela Figura 18. Todas as comparações de proteínas são feitas aplicando a heurística para identificação de similaridade proteica baseada em histograma de aminoácidos.

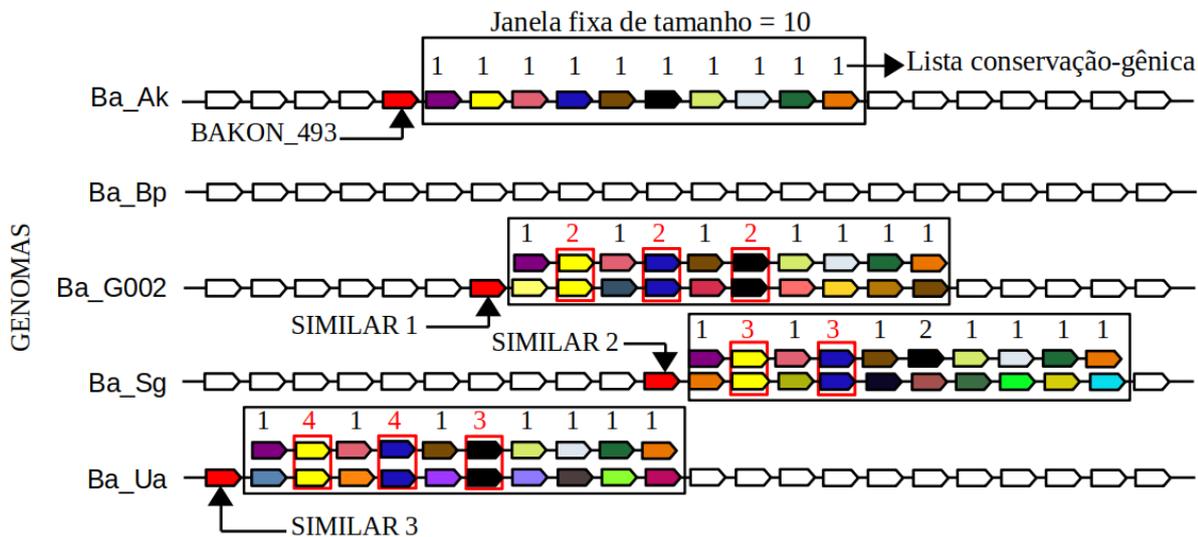


Figura 18 – Ilustração da expansão fixa com uma janela de tamanho igual a 10

Depois da comparação da vizinhança gênica da proteína BAKON_493 com a vizinhança das proteínas similares a ela, é verificado se alguma das 4 possíveis configurações de conservação gênica foi satisfeita dentro da janela. Essa verificação é feita na estrutura condicional *if/else-if* da linha 11 até 19 do Algoritmo 3. Primeiro é verificado se entre os 10 genes vizinhos subsequentes da proteína BAKON_493, constatou-se pelo menos 4 genes que se mantiveram conservados na vizinhança das proteínas similares a ela.

Para verificar o número de genes conservados é feita a contagem da quantidade de valores maiores que 1 existentes na lista *conservação-gênica* ilustrada na Figura 19. Caso exista pelo menos 4 genes conservados entre os 10 genes da janela, então cria-se arestas de PPI para todos os $n*(n-1)/2$ possíveis pares de proteínas da expansão. Vale mencionar que n é o número de genes envolvidos na expansão, isto é, desde a proteína BAKON_493 até a décima proteína subsequente a ela.

Janela de tamanho = 10; Conservação gênica requerida = 4; Conservação gênica constatada = 3.

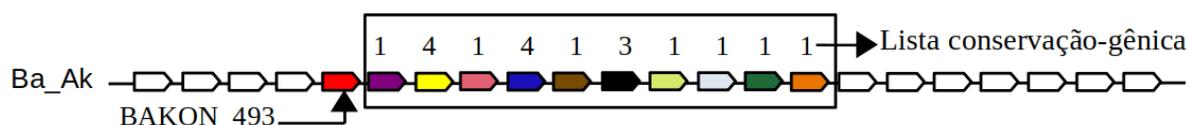


Figura 19 – Ilustração da primeira configuração de conservação gênica

Caso não exista 4 genes conservados na janela de 10, então verifica-se a próxima configuração de conservação gênica, que reduz o tamanho da janela e o número requerido de genes conservados de maneira condizente com a ilustração da Figura 20.

Janela de tamanho = 7; Conservação gênica requerida = 3; Conservação gênica constatada = 3.

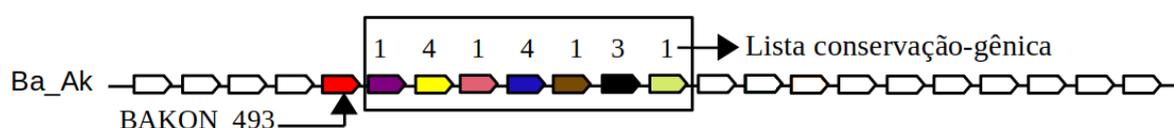


Figura 20 – Ilustração da segunda configuração de conservação gênica

Na segunda configuração de janela é verificado a seguinte condição: se entre os 7 genes vizinhos subsequentes da proteína BAKON_493, existirem pelo menos 3 genes conservados. Este é o caso ilustrado na Figura 20. Tendo essa condição sido satisfeita, então cria-se arestas de PPI para todos os possíveis pares de genes desse intervalo. Ou seja, desde a proteína BAKON_493 até a sétima proteína subsequente a ela. Caso não exista 3 genes conservados nessa redução de janela, então a terceira configuração de conservação gênica é verificada (2 genes conservados em uma janela de tamanho igual a 5). E assim até a quarta e última configuração (1 gene conservado em uma janela de tamanho igual a 3).

Quando uma dessas 4 configurações de conservação gênica é satisfeita, como é o caso ilustrado na Figura 20 (3 genes conservados numa janela de tamanho igual a 7), além de criar arestas de PPI para todos os pares de genes desse intervalo, também é atribuído um peso para essas arestas. Esse peso representa uma pontuação de confiança estimada para cada interação predita. Pontuação essa que depende do nível de conservação gênica das proteínas. Pares de proteínas vizinhas que coocorrem em mais genomas, recebem pesos maiores, e pares que coocorrem em menos genomas, recebem pesos menores. Os pesos das interações diretas (interações de pares de proteínas conservadas dentro da janela), são definidos da seguinte forma: calcula-se 65% da razão entre o número de genomas nos quais um dado par de proteínas está conservado e o número total de genomas incluídos na análise. Semelhantemente, o peso das interações indiretas (interações de pares de proteínas não conservadas da janela), é definido calculando 65% da razão entre o valor 1 e o número de genomas incluídos na análise. O valor 1 no cálculo dos pesos das interações indiretas,

é decorrente do fato desses pares de genes estarem em apenas 1 genoma da análise, não estando conservados em nenhum outro genoma.

Tomando de exemplo a conservação gênica da Figura 20, o valor 4 do gene representado pela cor amarelo (proteína BAKON_495), indica que esta proteína coocorre junto com a proteína BAKON_493 em 4 genomas. Isso implica em uma interação direta entre essas duas proteínas. Portanto, em concordância com o cálculo de pesos para interações diretas, o peso da aresta de interação entre as proteínas BAKON_493 e BAKON_495, seria $0.52 (4/5 * 0.65)$. O valor 4 neste cálculo indica o número de genomas nos quais as proteínas vizinhas BAKON_493 e BAKON_495 foram achadas conservadas na janela de expansão. O valor 5 representa o número de genomas incluídos na análise (exemplo da Figura 18). E o valor 0.65 indica a porcentagem destinada às interações proteicas preditas pelo método de vizinhança gênica conservada. Para exemplificar também o peso das interações indiretas, ou seja, interações entre genes não conservados da janela, considera-se a proteína BAKON_493 e a sua vizinha direita imediata, a proteína BAKON_494. Esta última é representada pelo gene de cor roxa na Figura 20. Mesmo que a proteína BAKON_494 não esteja conservada junto com a proteína BAKON_493 em outros genomas, acredita-se que essas duas também possuem algum nível de interação. A justificativa é que elas estão no mesmo *operon* (intervalo de genes onde foram identificados dois ou mais genes recorrentes). Portanto, cria-se também uma aresta de PPI entre as proteínas BAKON_493 e BAKON_494 atribuindo um valor menor ao peso dessa interação. De acordo com o cálculo definido para interações indiretas, o valor atribuído ao peso dessa aresta seria $0.13 (1/5 * 0.65)$. O peso das interações diretas ou indiretas dos outros pares de proteínas da janela de expansão, é calculado do mesmo modo que os exemplos dados neste parágrafo.

Em relação ao percentual utilizado para calcular o peso das arestas, deve-se considerar o seguinte: uma rede de interação gerada pelo GenPPI, além de representar as interações das proteínas de um organismo, também estima pontuações de confiança para as interações preditas. Essa pontuação pode ser de valores que variam entre 0 até 1. O motivo do uso de um percentual, é que uma mesma interação pode ser inferida por mais de uma das métricas implementadas, a saber, as métricas de vizinhança gênica conservada, fusão gênica, e perfil filogenético conservado. Por isso, utiliza-se um percentual no cálculo dos pesos das arestas de PPI inferidas pelas métricas abordadas. Assim, caso uma interação entre dois genes específicos, for inferida por mais de uma métrica, o peso dessa aresta não ultrapassará o valor máximo que deve ser igual a 1.0. Os valores padrões do programa para os percentuais das métricas de vizinhança gênica conservada, fusão gênica e perfil filogenético conservado são, respectivamente, 65%, 5% e 30%. Porém, esses percentuais podem ser alterados pelo o usuário.

Através da ilustração da Figura 18, pode-se observar que apenas a vizinhança gênica à direita da proteína BAKON_493 é analisada. A justificativa é que o Algoritmo 3 gera

interações redundantes quando se considera tanto a vizinhança à esquerda quanto a vizinhança à direita de um gene conservado. Para demonstrar essa redundância considera-se o exemplo da Figura 21.

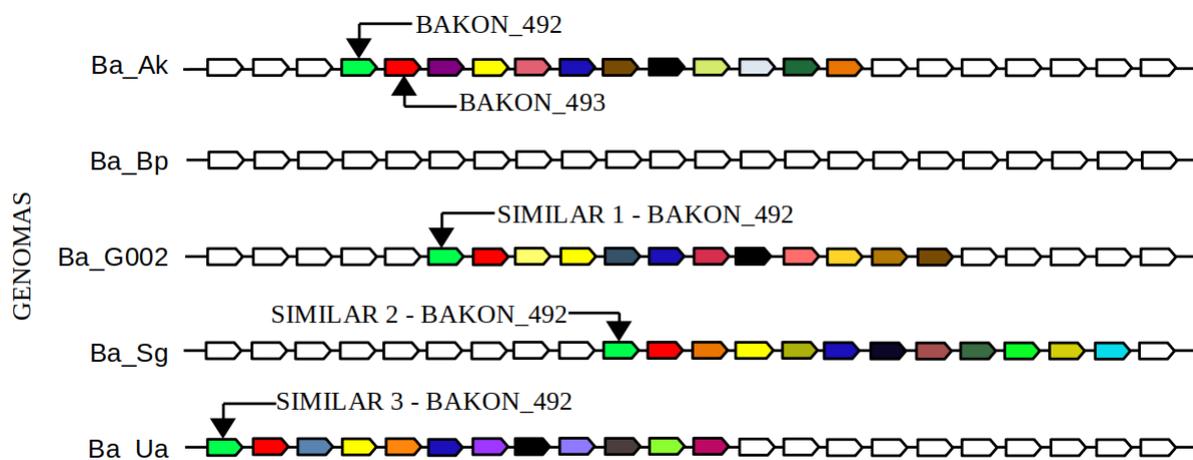


Figura 21 – Exemplificação de redundância na criação de arestas de PPI ao considerar tanto a vizinhança gênica à esquerda quanto à direita de uma proteína conservada

O laço *for all* da linha 1 do algoritmo de expansão fixa, é responsável por varrer a tabela-hash *pangenoma* verificando a existência de genes conservados na vizinhança das proteínas recorrentes, a fim de fazer inferências de PPIs. Considere o exemplo de conservação gênica ilustrado na Figura 21. Se antes da proteína BAKON_493 do genoma Ba_Ak, existisse outro gene que coocorre junto com ela em outros genomas, tal qual a proteína BAKON_492 representada pelo gene de cor verde, então essa proteína também estaria em um dos registros da tabela-hash *pangenoma*. Isso significa que em uma dada iteração do laço *for all* da linha 1, a vizinhança gênica à direita da proteína BAKON_492, seria analisada. Essa iteração produziria uma aresta de PPI que liga o gene de cor verde ao gene de cor vermelha: BAKON_492 - BAKON_493. Depois, na iteração subsequente do laço *for all* da linha 1, a vizinhança gênica da próxima proteína recorrente da tabela-hash *pangenoma*, seria analisada. Ou seja, a vizinhança da proteína BAKON_493. Assim, se além de considerar a vizinhança gênica à direita dessa proteína, também fosse considerada a vizinhança gênica à esquerda, isso produziria uma aresta de PPI redundante: BAKON_493 - BAKON_492. O motivo dessa redundância é que na iteração anterior do laço *for all* da linha 1, o inverso já teria sido inferido. Por essa razão, no Algoritmo 3 verifica-se apenas a vizinhança gênica à direita das proteínas conservadas.

Finalmente, após varrer toda tabela-hash *pangenoma* fazendo inferências de interações entre proteínas, uma tabela-hash denominada *ppi* guarda em seus registros as interações proteicas inferidas para os genomas incluídos em uma análise. Caso o conjunto de genomas analisado seja muito grande, algumas arestas de PPI, principalmente arestas de interações

indiretas, podem apresentar pesos muito baixos. Portanto, antes do algoritmo de expansão fixa ser finalizado, na linha 21 é feita uma normalização de pesos das arestas de PPI criadas para os genomas. Para tanto, considerando as arestas inferidas para um genoma específico, seus pesos são normalizados calculando 65% da razão entre o peso de uma dada aresta e o maior peso das interações desse genoma.

3.7.1.2 Expansão Dinâmica

Diferentemente da expansão fixa que usa uma janela de tamanho limitado, a expansão dinâmica não se limita a um intervalo fixo de genes, mas continua expandindo o tamanho da janela enquanto se identifica genes conservados. A configuração padrão para a expansão dinâmica estipula uma tolerância máxima de 2 genes não conservados, entre a identificação de um gene conservado e outro, para que a expansão da janela continue. A proposta da expansão dinâmica é identificar trechos maiores de genes conservados, não interrompidos por uma quantidade muito grande de genes não conservados. O Algoritmo 4 demonstra o funcionamento da expansão dinâmica.

Algoritmo 4 Expansão Dinâmica

Entrada: *pangenoma*, *percentual-vgc*, *ws*, *d-limite*, *qtd-amin*.

Saída: tabela-hash *ppi*.

```

1: for all proteína ∈ pangenoma do
2:   pivô-1 ← proteína.localidade
3:   lista conservação-gênica ← (1 1 1)
4:   for all pivô-2 ∈ proteína.similares do
5:     pos ← 0
6:     repeat
7:       gene-conservado ← false
8:       for i = 1 to ws do
9:         if heurística(pivô-1+(pos+i), pivô-2+(pos+i), d-limite, qtd-amin) then
10:          gene-conservado ← true
11:          incrementa(conservação-gênica[pos+(i-1)])
12:          expandir(conservação-gênica)
13:          pos ← + i
14:          i ← ws
15:        end if
16:      end for
17:    until gene-conservado = false
18:  end for
19:  if conservação-gênica then
20:    criar-arestas-ppi(conservação-gênica, percentual-vgc, ppi)
21:  end if
22: end for
23: normalizar-pesos(ppi)
24: return tabela-hash ppi

```

Igualmente ao algoritmo de expansão fixa, o de expansão dinâmica varre a tabela-hash *pangenoma* aplicando uma sequência de ações a cada iteração. O objetivo é fazer inferências de PPIs baseadas na constatação e eventos de vizinha gênica conservada. Para demonstrar o funcionamento da expansão dinâmica, considera-se novamente uma iteração do laço *for all* da linha 1 aplicada à proteína conservada BAKON_493 da Figura 17.

Inicialmente, é necessário verificar se a proteína conservada BAKON_493 faz parte de um intervalo recorrente de genes. Para tanto, da linha 4 até a linha 18 do Algoritmo 4 é feita a comparação da vizinhança gênica da proteína BAKON_493 (variável *pivô-1*) com a vizinhança das proteínas similares a ela (variável *pivô-2*). Nessa comparação, conta-se em quantos genomas houve conservação da vizinhança gênica da proteína BAKON_493. Todas as comparações de proteínas são feitas aplicando a heurística baseada em histograma de aminoácidos. Diferente da expansão fixa que verifica um intervalo limitado da vizinhança gênica dessas proteínas, a expansão dinâmica continua expandindo a janela progressivamente até que a qualidade de conservação diminua. Considerando a configuração padrão do programa para o método de expansão dinâmica, a expansão da janela somente é encerrada quando não se identificar nenhum outro gene conservado em um intervalo de 3 genes subsequentes. Esse processo é ilustrado pela Figura 22.

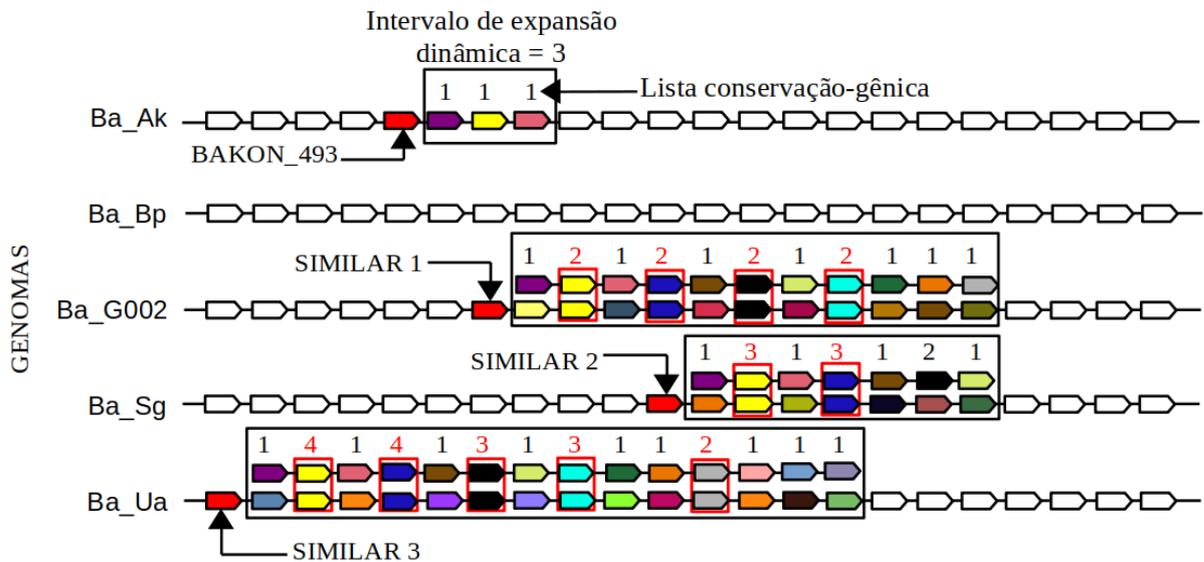


Figura 22 – Ilustração da expansão dinâmica

A expansão dinâmica é uma verificação progressiva que expande a janela de 3 em 3 enquanto se identifica proteínas conservadas. Primeiro é feita a comparação da vizinhança gênica da proteína BAKON_493 com a vizinhança da proteína similar 1. A expansão da janela continua progressivamente até que 3 genes consecutivos sejam averiguados sem identificar nenhum novo gene conservado. Depois, o mesmo é feito com as proteínas similares 2 e 3 ilustradas na Figura 22. A medida que são identificados genes conservados na vizinhança da proteína BAKON_493, a lista de conservação gênica é incrementada. O

objetivo é contar em quantos genomas esses genes estão conservados. Ao finalizar todas as comparações, a lista de conservação gênica possui a configuração final de valores (dados fictícios) ilustrada na Figura 23.

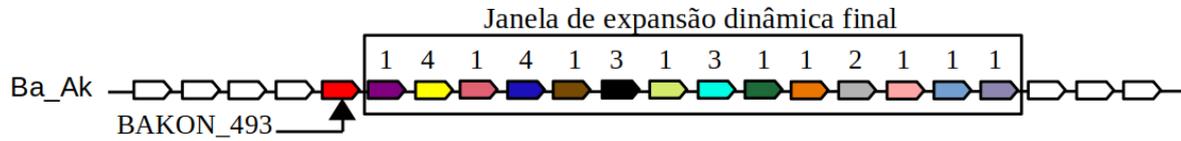


Figura 23 – Ilustração da janela de expansão dinâmica final

Apenas os genes que se repetiram em outros genomas, juntos com a proteína conservada BAKON_493, é que possuem valores maiores que 1 na lista de conservação gênica gerada pela expansão dinâmica. Os valores dessa lista representam em quantos genomas dentre o total de 5 analisados, um gene está conservado. A Figura 23 ilustra a proposta da expansão dinâmica, ou seja, identificar trechos maiores de genes conservados, não interrompidos por uma quantidade muito grande de genes não conservados. Visando esse objetivo a expansão dinâmica é interrompida ao atingir seu critério de parada: averiguar 3 genes subsequentes sem identificar nenhum novo gene conservado. Vale mencionar que esse critério é a configuração padrão do programa. Porém, essa configuração pode assumir qualquer valor informado pelo usuário através do parâmetro *-ws* do GenPPI.

Após o processo de expansão ilustrado na Figura 22, na linha 19 do Algoritmo 4 é verificado se genes conservados foram identificados. Caso sim, então cria-se arestas de PPI para todos possíveis pares de proteínas do intervalo recorrente de genes identificado. Ou seja, desde a proteína BAKON_493 até a última proteína alcançada pela expansão dinâmica. Sobre o cálculo de pesos dessas arestas dispensa-se comentários, pois esse cálculo é idêntico ao cálculo de pesos já explicado na subseção anterior referente à expansão fixa.

Finalmente, o algoritmo de predição de PPI pelo método de expansão dinâmica, termina depois de averiguar a vizinhança gênica de todas as proteínas conservadas da tabela-hash *pangenoma*. Tal qual o algoritmo de expansão fixa, a saída do algoritmo de expansão dinâmica é uma tabela-hash denominada *ppi* contendo as interações previstas para os genomas incluídos em uma análise. Também vale mencionar que a mesma normalização de pesos aplicada na expansão fixa, também é aplicada na linha 23 do algoritmo de expansão dinâmica, antes do mesmo ser finalizado.

3.7.2 Implementação da Predição de PPI Pelo Método de Fusão Gênica

Em conformidade com a Subseção 2.2.5.2 do capítulo de fundamentação teórica, uma fusão de genes se refere a um evento em que dois genes individuais de um dado organismo

se fundem numa sequência única formando um novo gene em outro organismo. Esse método pressupõe que a fusão entre dois genes implique em uma interação física ou funcional de seus produtos proteicos. Para inferir PPIs baseadas em eventos de fusão gênica, foi implementado o Algoritmo 5 apresentado a seguir e comentado na sequência.

Algoritmo 5 Fusão Gênica

Entrada: tabela-hash *genomas*, percentual-fg, *d-limite*, *qtd-amin*.

Saída: tabela-hash *ppi*.

```

1: for all  $genoma_i \in genomas$  do
2:   for all  $proteína_i \in genoma_i$  do
3:      $fusão \leftarrow soma-histogramas(proteína_i, proteína_{i+1})$ 
4:     for all  $genoma_j \in genomas$  do
5:       if not( $genoma_i = genoma_j$ ) then
6:         for all  $proteína_j \in genoma_j$  do
7:           if  $heurística(fusão, proteína_j, d-limite, qtd-amin)$  then
8:              $criar-arestas-ppi(genoma_i.nome, proteína_i, proteína_{i+1}, ppi)$ 
9:           end if
10:        end for
11:      end if
12:    end for
13:  end for
14: end for
15: return tabela-hash ppi

```

Este algoritmo é muito simples, pois basicamente o que ele faz é averiguar se dois genes (A e B) vizinhos imediatos na sequência gênica de um genoma específico, se fundem numa sequência única formando um novo gene em outro organismo. Segundo ilustra a Figura 24, se tal evento for constatado, uma interação é inferida para os genes A e B.

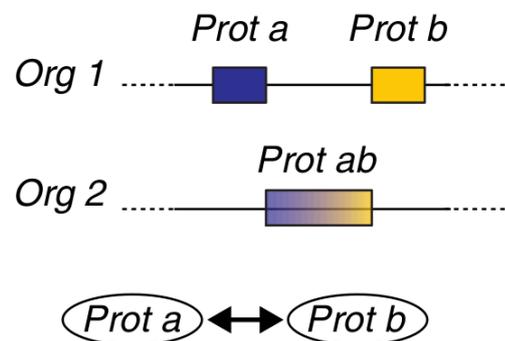


Figura 24 – Interação proteica por fusão gênica

Fonte: (VALENCIA; PAZOS, 2002)

No GenPPI, a fusão entre duas proteínas A e B é representada pela soma de seus histogramas de aminoácidos. Um exemplo pode ser visto na Tabela 4.

Tabela 4 – Fusão gênica via soma dos histogramas de aminoácidos de duas proteínas (A e B).

Aminoácidos:	A	R	N	D	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V	U	O	X	B	Z	J
Histograma A:	20	8	7	9	1	7	6	14	5	11	18	12	3	11	13	28	34	2	4	15	0	0	0	0	0	0
Histograma B:	45	16	14	19	3	14	18	38	7	20	45	6	10	12	37	27	35	7	13	29	0	0	0	0	0	0
Fusão:	65	24	21	28	4	21	24	52	12	31	63	18	13	23	50	55	69	9	17	44	0	0	0	0	0	0

No algoritmo de fusão gênica os laços *for all* das linhas 1 e 2, respectivamente, iteram entre todos os genomas de uma análise e entre todas as suas proteínas. O objetivo é formar pares de proteínas vizinhas subsequentes em cada genoma. Na linha 3, de maneira condizente com o exemplo da Tabela 4, soma-se (fusão) os histogramas de aminoácidos de um dado par de proteínas (A e B) de um genoma específico gerando um novo gene. Depois, da linha 4 até 12 esse novo gene fundido é comparado com todos os genes de outros genomas, aplicando a heurística do GenPPI. Um evento de fusão gênica é identificado quando na linha 7, a heurística identificar similaridade proteica entre o gene fundido e o *j*-ésimo gene da comparação. Isso gera uma aresta de interação para o par de proteínas (A e B) utilizados para gerar o gene fundido.

Como dito anteriormente, o peso de uma aresta de interação pode variar entre 0 até 1 representando um nível de confiança atribuído a essa interação predita. Sobre o peso das arestas de PPI geradas por fusão gênica, considerando as configurações padrões do programa, todas recebem o mesmo peso, o valor 0,05 que corresponde a 5% do peso máximo. A justificativa pelas interações inferidas por eventos de fusão gênica, terem apenas 5% do peso máximo igual a 1, é que esses eventos são mais raros de acontecer (PAN-CHENKO; PRZYTZYCKA, 2010). Assim, os mesmos geram uma quantidade pequena de interações, na casa das unidades, enquanto eventos de vizinhança gênica conservada e perfil filogenético conservado, geram na casa das centenas e milhares de interações.

O Algoritmo 5 termina após finalizada a busca de eventos de fusão gênica, para todo par de proteínas vizinhas subsequentes presentes nos genomas do conjunto analisado. A saída desse algoritmo é uma tabela-hash de nome *ppi* utilizada para armazenar as interações inferidas para todos os genomas.

3.7.3 Implementação da Predição de PPI Pelo Método de Perfil Filogenético Conservado

O método de inferência de interações proteicas por eventos de perfil filogenético conservado, é baseado na hipótese de que todas as proteínas funcionalmente ligadas (em interação) tendem a ser preservadas ou eliminadas juntas em uma nova espécie durante a evolução. Ou seja, se duas proteínas são necessárias para a realização de uma determinada função biológica, um organismo precisa possuir ambas as proteínas caso essa função seja necessária, enquanto ambas não fazem falta caso essa função não seja necessária. Isso indica uma interação entre essas duas proteínas (PELLEGRINI et al., 1999). Baseado nessa hipótese, esse método de predição de PPIs faz um mapeamento da presença ou ausência das proteínas de um organismo de consulta em relação a um grupo de organismos de referência. A partir desse mapeamento são feitas inferências de interações entre proteínas que ocorrem juntas e estão ausentes juntas nos mesmos organismos de referência. Ou seja, entre proteínas com o mesmo perfil filogenético.

Para fazer predições de PPIs com esse método, primeiro se constrói o perfil filogenético das proteínas de um organismo de interesse (organismo de consulta). Esse perfil é um registro da presença ou ausência de proteínas similares em um conjunto de organismos de referência. Conforme demonstrado na Subseção 2.2.5.2 do capítulo de fundamentação teórica, o perfil filogenético de uma dada proteína é representado por uma cadeia de bits de comprimento n . O valor n é referente ao número de organismos de referência. Cada bit da cadeia corresponde a um organismo: se um dado bit é 1, isso indica que a proteína em questão possui um gene similar nesse organismo, se é 0 não possui um gene similar. Depois, proteínas que têm cadeias de bits (perfis filogenéticos) iguais ou semelhantes, são agrupadas e consideradas interagindo umas com as outras.

O Algoritmo 6 demonstra de maneira simplificada, a implementação de inferência de interações proteicas pelo método de perfil filogenético conservado.

Algoritmo 6 Perfil Filogenético Conservado

Entrada: tabela-hash *genomas*, *percentual-fg*, *dif-tolerada*, *d-limite*, *qtd-amin*.

Saída: tabela-hash *ppi*.

```

1: tabela-hash perfis-filogeneticos  $\leftarrow$  vazio
2: gerar-perfis(perfis-filogeneticos, genomas, d-limite, qtd-amin)
3: for all genomai  $\in$  perfis-filogeneticos do
4:   agrupamentos  $\leftarrow$  agrupar(genomai, dif-tolerada)
5:   for all grupo  $\in$  agrupamentos do
6:     criar-arestas-ppi(grupo, ppi)
7:   end for
8: end for
9: return tabela-hash ppi

```

Inicialmente, na linha 1 do algoritmo de inferência de PPIs pelo método de perfil filogenético conservado, uma tabela-hash denominada *perfis-filogeneticos* é inicializada vazia. Essa tabela-hash é utilizada para armazenar os perfis filogenéticos das proteínas conservadas dos genomas. A função *gerar-perfis* da linha 2 varre a tabela-hash *genomas* montando o perfil filogenético das proteínas. Para montar o perfil de uma dada proteína, a função *gerar-perfis* considera como organismos de referência, todos os genomas incluídos em uma análise com exceção do genoma ao qual essa proteína pertence (genoma de consulta). Assim, supondo que em uma análise foram incluídos 6 genomas, o perfil filogenético de uma proteína do genoma de consulta, seria uma cadeia de 5 bits. Cada bit representando a presença ou ausência dessa proteína em um dos 5 genomas de referência, conforme ilustra a Figura 25. Posteriormente, as proteínas com os seus devidos perfis montados, são inseridas na tabela-hash *perfis-filogeneticos*. Nessa tabela, um registro corresponde a uma lista com os perfis das proteínas de um dado genoma.

Depois, no laço *for all* da linha 3 o primeiro passo (linha 4) é a geração de uma lista para um *genoma_i*, contendo grupos de proteínas com perfis filogenéticos idênticos.

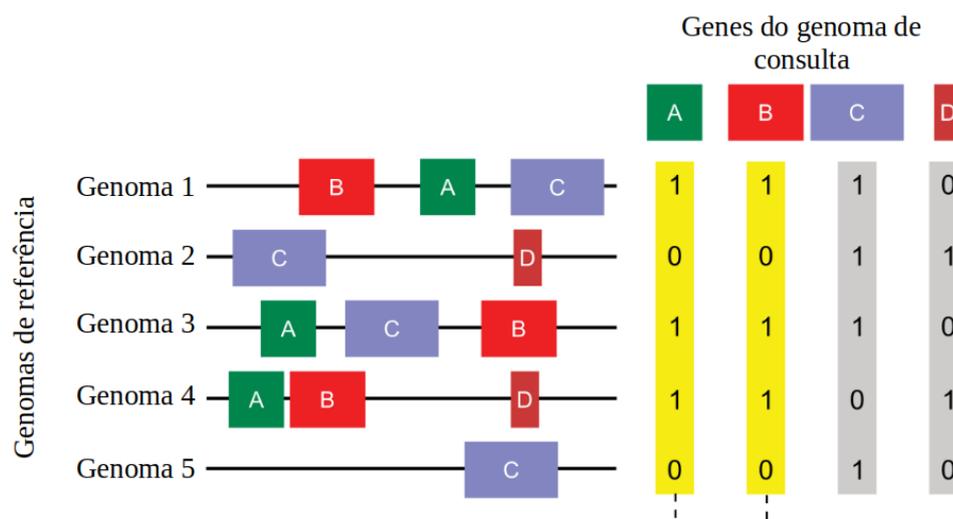


Figura 25 – Perfil filogenético conservado

Fonte: Adaptado de (RAMAN, 2010)

Na sequência, o laço *for all* da linha 5 varre essa lista criando arestas de PPI para os $n*(n-1)/2$ possíveis pares de proteínas dos grupos formados em um $genoma_i$.

Quanto ao o peso dessas arestas, considera-se a seguinte regra: as arestas dos pares de proteínas do grupo mais populoso, recebem peso igual a 0.30, valor que corresponde a 30% do peso máximo (1.0). O peso das arestas dos pares de proteínas dos outros grupos, é definido calculando 30% da razão entre o número de proteínas do grupo em questão e o número de proteínas do grupo mais populoso.

O Algoritmo 6 termina após todas as iterações do laço *for all* da linha 3, ou seja, após criar arestas de PPI para todos os possíveis pares de proteínas dos agrupamentos gerados para os genomas da análise. A saída desse algoritmo também é a uma tabela-hash com o nome *ppi*, utilizada para armazenar as interações previstas para os genomas analisados.

Conforme pode ser observado na linha 4 do Algoritmo 6, a função *agrupar* possui um parâmetro chamado *dif-tolerada*. Esse parâmetro determina se os agrupamentos serão formados somente de proteínas com perfis filogenéticos idênticos (*dif-tolerada* = 0), ou também de proteínas com perfis semelhantes (*dif-tolerada* > 0). Dois perfis filogenéticos são considerados semelhantes se e somente se a diferença de suas cadeias de bits não for maior que a quantidade de bits de diferença tolerada pelo parâmetro *dif-tolerada*. Assim, se a diferença tolerada for igual a 1 por exemplo, proteínas cujos perfis filogenéticos diferem em no máximo 1 bit, serão consideradas proteínas com perfis semelhantes. Este é o caso do exemplo ilustrado pela Figura 26.

Por padrão, as predições pelo método de perfil filogenético conservado, são feitas somente para pares de proteínas com perfis filogenéticos idênticos. Caso o usuário também queira considerar proteínas com perfis semelhantes como interagindo entre si, a diferença

dif-tolerada = 1

Proteína A: 1 0 1 1 0 1 1 0 1 1
 Proteína B: 1 0 1 1 0 0 1 0 1 1

Figura 26 – Ilustração de perfis semelhantes

de perfis tolerada deverá ser informada no momento da execução do programa.

3.8 Encerramento

Após a predição de PPI por meio das métricas de vizinhança gênica conservada, fusão gênica e perfil filogenético conservado, o programa finaliza suas análises gerando redes de interação de proteínas para todos os genomas do conjunto analisado. Essas redes são todas as interações par-a-par de proteínas, preditas pelas 3 métricas para todos genomas. As interações de cada organismo (genoma) são salvas em um arquivo com a extensão *.dot* para análise posterior no programa GEPHI (programa para análise topológica de redes e grafos). Para fins de demonstração, a Figura 27 apresenta um pequeno trecho de um arquivo *.dot* gerado pelo GenPPI.

```
"BAKON_493" -- "BAKON_494" [WEIGHT = 0.13];
"BAKON_493" -- "BAKON_495" [WEIGHT = 0.52];
"BAKON_493" -- "BAKON_496" [WEIGHT = 0.13];
"BAKON_493" -- "BAKON_497" [WEIGHT = 0.52];
"BAKON_493" -- "BAKON_498" [WEIGHT = 0.13];
"BAKON_493" -- "BAKON_499" [WEIGHT = 0.39];
"BAKON_493" -- "BAKON_500" [WEIGHT = 0.13];
"BAKON_494" -- "BAKON_495" [WEIGHT = 0.13];
"BAKON_494" -- "BAKON_496" [WEIGHT = 0.13];
"BAKON_494" -- "BAKON_497" [WEIGHT = 0.13];
"BAKON_494" -- "BAKON_498" [WEIGHT = 0.13];
"BAKON_494" -- "BAKON_499" [WEIGHT = 0.13];
"BAKON_494" -- "BAKON_500" [WEIGHT = 0.13];
"BAKON_495" -- "BAKON_496" [WEIGHT = 0.13];
"BAKON_495" -- "BAKON_497" [WEIGHT = 0.52];
```

Figura 27 – Trecho de um arquivo de interação proteína-proteína gerado após uma execução do GenPPI para analisar 5 linhagens da bactéria *Buchnera Aphidicola*

A Figura 27 apresenta um trecho do arquivo *.dot* de interação proteína-proteína gerado para a cepa *Acyrtosiphon kondoi* da bactéria *Buchnera Aphidicola*, após uma aplicação do programa sob 5 linhagens dessa bactéria. Uma linha do arquivo representa uma aresta de interação da rede de PPI gerada. Por exemplo, a segunda linha da Figura 27 indica uma interação entre as proteínas BAKON_493 e BAKON_495, com peso igual a 0.52. Vale lembrar que o peso das arestas indica uma pontuação de confiança estimada para as interações preditas, pontuação essa que pode variar entre 0 até 1.

3.9 Complexidade Algorítmica do GenPPI

No GenPPI o algoritmo para análise de vizinhança gênica conservada com expansão dinâmica (Algoritmo 4), é o que demanda maior tempo para ser processado. Portanto, a complexidade temporal do GenPPI é calculada por meio desse algoritmo. O mesmo consiste em quatro comandos de repetição aninhados, sendo que um deles não possui uma variável específica cujo limite estipula o fim da execução de repetições. A seguir, as variáveis que influenciam o algoritmo para análise de vizinhança gênica conservada em modo de expansão dinâmica, são representadas por letras gregas com exceção do termo ws que é um parâmetro do programa.

- ν = número médio de proteínas contidas nos genomas analisados;
- μ = número de genomas da análise;
- σ = similaridade média entre os genomas ao nível proteico;
- ws = parâmetro do GenPPI que representa o tamanho da janela e passo incremental para expansão dinâmica;
- ρ = uma constante específica para cada conjunto de genomas.

O primeiro loop no Algoritmo 4 (linha 1) garante inspecionar todas as proteínas conservadas identificadas na geração do pangenoma. O número de proteínas para um conjunto de genomas, depende do número médio de proteínas por genoma (ν) multiplicado pelo número de genomas (μ). Na linha 2, uma proteína conservada se torna a principal (pivô-1) para as análises de vizinhança gênica conservada. Como a heurística do GenPPI (Algoritmo 2) trabalha em pares de proteínas, na linha 4 do Algoritmo 4, o segundo loop seleciona uma proteína homóloga à proteína pivô-1 (pivô-2), a fim de verificar se há genes conservados na vizinhança gênica desse par de proteínas em seus genomas. Na linha 4, a quantidade de proteínas homólogas à pivô-1, depende de μ multiplicado pela similaridade média entre os genomas ao nível proteico (σ). Na linha 8, o parâmetro do GenPPI ws , definido em tempo de execução, define o valor da variável ws (tamanho da janela de análises de vizinhança gênica conservada), especificando o limite de execução do quarto loop interno. Um valor ws significa um passo progressivo que expande a janela enquanto se constata conservação gênica na vizinhança das proteínas pivô-1 e pivô-2. Por exemplo, suponha que ws seja igual a 3 e que o Algoritmo 4 encontre uma conservação gênica em alguma posição desse intervalo de janela. Nesse caso, ele registra essa conservação gênica e se prepara para verificar os próximos 3 genes subsequentes à posição da janela onde se constatou a conservação gênica. Conseguimos esse passo a passo progressivo incrementando monotonicamente a variável pos . Na linha 6, existe um loop cuja condição de término (linha 17) é responsável pela incerteza sobre a complexidade temporal

do Algoritmo 4. O término desse algoritmo para um par de genomas acontecerá quando uma janela atual expandida ($ws_{current} = (pos / ws_{initial}) \cdot ws_{initial}$) não constatar mais uma quantidade mínima de proteínas conservadas ao realizar um passo de expansão. Não podemos inferir uma fórmula exata em relação ao número de iterações do loop da linha 6. No entanto, podemos tentar tabulá-la incrementando o valor ws para um conjunto de genomas. Chamamos essa variável de ρ . Afinal, os dois loops mais internos do Algoritmo 4 têm uma complexidade proporcional a ρ multiplicada por ws .

Dado um par de valores conhecidos de ρ , ν e σ , poderíamos tentar aproximar ρ para um determinado conjunto de genomas. ρ tem valores que são proporcionais a ν e σ . Poderíamos, por exemplo, estabelecer um sistema quadrático de equações lineares (1) que permitiria aproximar valores de ρ . Na equação 1, suponha que os dados dos genomas conhecidos sejam de *Corynebacterium* (cp) e *Staphylococcus* (st).

$$\begin{aligned} K_{\sigma} \cdot \sigma_{cp} + K_{\nu} \cdot \nu_{cp} &= K_{cp} \\ K_{\sigma} \cdot \sigma_{st} + K_{\nu} \cdot \nu_{st} &= K_{st} \end{aligned} \quad (1)$$

Uma vez estimadas as constantes da equação (1), pode ser possível estimar o valor de ρ (ν , σ) com a equação 2 para um genoma (g):

$$\rho_g(\nu_g, \sigma_g) = K_{\nu} \cdot \nu_g + K_{\sigma} \cdot \sigma_g \quad (2)$$

A complexidade temporal do GenPPI foi inferida pela presença das relações abaixo por linha no Algoritmo 4:

- $\nu \cdot \mu =$ linha 1;
- $\mu \cdot \sigma =$ linha 4;
- $\rho \cdot ws =$ linhas 6 e 8.

Finalmente, a quantidade de comparações feitas entre as proteínas de um conjunto de genomas, considerando o algoritmo de expansão dinâmica para análise de vizinhança gênica conservada (Algoritmo 4), é estimada com a equação 3.

$$O(\nu \cdot \mu^2 \cdot \sigma \cdot (\rho(\nu, \sigma)) \cdot ws) \quad (3)$$

Enfatizamos que estas são estimativas. Por exemplo, para *Staphylococcus*, com ws variando entre 5, 6 e 7 foram gastas 68, 96 e 102 horas, respectivamente. Não há garantia que haverá um aumento monotônico da complexidade de um valor de ws para outro superior, ou seja, que o fator de repetição ρ será sempre mantido. Não há uma diferença constante entre as execuções com $ws = 5$ para $ws = 6$ (28 horas) e de $ws = 6$ para $ws = 7$ (sete horas). Isso caracteriza uma incerteza na quantidade de execuções desse algoritmo. Porém, podemos contar com um valor médio para essa inflação de execuções entre diferentes valores de ws .

3.10 Otimização de Código

Para resolver problemas de estouro de memória foi necessário o emprego de funções de otimização de código do Common Lisp. Quando não utilizamos definições de dados no Lisp temos objetos de tamanho padronizado e genéricos. A consequência é um maior tempo de processamento e maior uso de memória. Considerando que a quantidade de dados a serem processados pode ser muito expressiva, fez-se necessário a realização de uma otimização de código visando evitar problemas de estouro de memória.

Há diversas dicas de otimização de código no Common Lisp. Pode-se otimizar o compilador para gerar códigos focando em velocidade, alocação de memória, debug e segurança. Essa configuração é muito simples, a princípio. A ideia básica é colocar uma cláusula `declaim` na primeira linha do código Lisp do programa: **(`declaim (optimize (speed 3) (space 0) (debug 0) (safety 0))`)**. Essa cláusula cria códigos de máquina mais compactos que executam mais rápido do que se houvessem códigos para facilitar o debug e a segurança (LISPCOOKBOOK, 2021). Conseqüentemente, o consumo de memória RAM é menor também, pois não seria mais utilizado uma área de memória genérica. Entretanto, o emprego apenas dessa cláusula não foi suficiente para evitar problemas de estouro de memória. A solução foi o uso dessa cláusula aliado à tipagem de variáveis de entrada e saída de diversas funções implementadas no programa.

Visando demonstrar a eficácia dos recursos de otimização de código do Common Lisp, foi feito um experimento com cinco variações de uma função implementada para gerar a sequência de Fibonacci, dado um número n recebido como parâmetro. Essas variações se diferem apenas na quantidade de recursos de otimização de código empregada. As funções Lisp dessas 5 variações são apresentadas na Tabela 5.

Tabela 5 – Variações de recursos de otimização de código Lisp, aplicados a uma função que gera a sequência de Fibonacci

Variações	Implementação	Explicação
Fib-1	<pre>(defun fib1(n) (if (< n 2) 1 (+ (fib1 (- n 1)) (fib1 (- n 2)))))</pre>	Sem nenhuma otimização de código.
Fib-2	<pre>(defun fib2(n) (declare (type fixnum n)) (if (< n 2) 1 (+ (fib2 (- n 1)) (fib2 (- n 2)))))</pre>	Tipagem apenas da variável interna.
Fib-3	<pre>(declaim (ftype (function (fixnum) fixnum) fib3)) (defun fib3(n) (if (< n 2) 1 (+ (fib3 (- n 1)) (fib3 (- n 2)))))</pre>	Tipagem apenas dos parâmetros da função.
Fib-4	<pre>(declaim (ftype (function (fixnum) fixnum) fib4)) (defun fib4(n) (declare (type fixnum n)) (if (< n 2) 1 (+ (fib4 (- n 1)) (fib4 (- n 2)))))</pre>	Tipagem da variável interna e dos parâmetros da função.
Fib-5	<pre>(declaim (optimize (speed 3) (debug 0) (safety 0))) (declaim (ftype (function (fixnum) fixnum) fib5)) (defun fib5(n) (declare (type fixnum n)) (if (< n 2) 1 (+ (fib5 (- n 1)) (fib5 (- n 2)))))</pre>	Além da tipagem da variável interna e dos parâmetros da função, foi feita a inserção de uma cláusula <i>declaim</i> na primeira linha do código visando otimizar o tempo de execução.

Tabela 6 – Quantidades de ciclos executados pelo processador nas 5 variações de implementações da função de Fibonacci

Execução	Ciclos Fib-1	Ciclos Fib-2	Ciclos Fib-3	Ciclos Fib-4	Ciclos Fib-5
1	14596731174	12464360592	9345900992	9189243671	6357881346
2	14860394220	11560988112	9296207487	9161572956	6141031770
3	14670395664	12545230059	9797814156	9257460642	6390454659
4	14611385442	12156575982	10148029836	9532117962	6199360242
5	14646502644	11966047467	9809345916	9222131538	6303155517
6	14859419541	12525107457	10027717002	9330417960	6264288216
7	14686419744	11472073269	9932761425	9196687626	6140974743
8	14870830254	12223041885	10592751429	9200514630	6274148688
9	15117093498	11834069565	9742087635	9267605307	6230985348
10	15513215532	12557735259	10166266857	9173630826	6141747294
11	15177548676	12140939856	9637060320	9479031966	6110888985
12	15218602365	12039637761	10276799403	9267070140	6134236344
13	15657554514	12542955777	14739350886	9260320689	6095752458
14	15755067348	11857522134	11343293258	10079226975	6216954888
15	15697814280	12164227623	9889778088	9639837363	6289620750
16	15498468456	12424130964	10157851479	10029754731	6210609075
17	15306818616	11880040902	9745059609	9551755092	6177635157
18	15569027712	12346267323	10151008233	10073614554	6241731429
19	15682134924	12233715300	9753437922	9448468083	6135529566
20	15770517681	17269736144	10517057910	9267477513	6210486549
21	15808564833	12299247417	14771784779	10051639230	6114071376
Média	15248888797.2	12401964512.8	10524773181.5	9474516789.15	6201183152.7

O experimento realizado consistiu na contagem de ciclos executados pelo processador para processar separadamente as 5 variações da função de Fibonacci para $n = 42$. A contagem de ciclos de máquina foi feita utilizando a função *time* do Common Lisp. Cada variação implementada foi executada 21 vezes através de um laço de repetição. Por final, visando comparar o desempenho das variações implementadas, calculou-se a média da quantidade de ciclos executados pelo processador nas 21 execuções das funções. Os resultados desses testes estão apresentados na Tabela 6. Nessa tabela pode-se observar que a média de ciclos executados pelo processador nas cinco variações da função de Fibonacci, está decrescente de Fib-1 para Fib-5. Tal decréscimo representa uma queda de 59,33% na média de ciclos da primeira para a quinta variação. Ou seja, quanto mais recursos de otimização de código foram empregados, maior foi a velocidade de processamento da mesma função. Consequentemente, o consumo de memória também é menor, pois não se utiliza mais uma área de memória genérica. Isso porque foi feita a tipagem de variáveis das funções conforme demonstrado na Tabela 5. Assim, os recursos de otimização de código Lisp apresentados na Tabela 5, aplicados ao longo de mais de 5 mil linhas de código do GenPPI, foram suficientes para resolver problemas de estouro de memória.

Experimentos e Análise dos Resultados

Este capítulo irá discorrer sobre os experimentos e análises de resultados que foram realizados para validar a hipótese principal. Esta é referente à possibilidade de fazer previsões *ab initio* de redes de PPI biologicamente corretas e confiáveis através da ferramenta computacional proposta, além de realizar tal tarefa em tempo aceitável.

4.1 Métodos de Avaliação

Nesta seção são descritos os métodos utilizados para validar a hipótese principal, bem como métodos adotados para avaliar a otimização de código realizada e a heurística proposta. Também é falado sobre as bases de dados utilizadas e os trabalhos com os quais a proposta deste é comparada.

4.1.1 Método Para Avaliação de Resultados da Otimização de Código Realizada

Para resolver problemas de estouro de memória, foram utilizados recursos de otimização de código do Lisp. Visando demonstrar o impacto de tal otimização no consumo de memória RAM, utilizou-se uma base de dados constituída por 50 genomas de *Corynebacterium pseudotuberculosis*. Estes foram processados por duas versões do programa, uma otimizada e a outra não. Os resultados são apresentados em três categorias: consumo de memória RAM, quantidade de ciclos executados pelo processador e o tempo gasto. Assim, além do impacto da otimização de código em termos de uso de memória RAM, demonstra-se também o tempo gasto para processar um conjunto expressivo de genomas. O objetivo é concluir se a solução desenvolvida realiza previsões *ab initio* em um tempo aceitável ou não. Para checar a alocação de memória, a quantidade de ciclos executados pelo processador e o tempo de processamento das versões sem otimização e com otimização de código, foram utilizadas as funções *room* e *time* do Common Lisp.

4.1.2 Método Para Avaliação da Heurística Proposta

Com o objetivo de encontrar valores para os parâmetros da heurística do GenPPI, que garantissem um percentual satisfatório de identidade mínima (90%) para todo par de proteínas classificado similar pela heurística, foram feitos vários testes frente ao algoritmo exato de alinhamento de sequências Needleman-Wunsch (NEEDLEMAN; WUNSCH, 1970). Para tanto, utilizou-se uma base de dados constituída por 3988 sequências de aminoácidos de proteínas oriundas do genoma de *Mycobacterium tuberculosis H37Rv*. Esses testes foram feitos combinando diferentes valores para os parâmetros da heurística, com o objetivo de encontrar combinações que garantissem uma exatidão satisfatória na classificação de similaridade proteica. Tal garantia foi dada pelo algoritmo exato Needleman-Wunsch utilizado para estimar um percentual de identidade para cada par de proteínas classificado similar pela heurística.

Além disso, o desempenho da heurística proposta foi comparado com o desempenho do BLASTp (ALTSCHUL et al., 1990), o principal algoritmo heurístico do estado da arte para comparação de sequências de proteínas. Verificou-se qual dos dois algoritmos heurísticos (GenPPI vs BLASTp) se aproxima mais do algoritmo exato Needleman-Wunsch em termos de exatidão na comparação par-a-par de proteínas. O experimento realizado consistiu na utilização dos três algoritmos para identificar pares de proteínas similares em uma cepa de *Mycobacterium tuberculosis* com 4026 proteínas. Assumimos como um par similar, duas proteínas com no mínimo 90% de identidade de sequência. O método para comparação de desempenho empregado faz uso de medidas como: exatidão na comparação par-a-par de proteínas, tempo de processamento e uso de memória RAM. Através da ferramenta DB Browser for SQLite, foi criado um pequeno banco de dados com três tabelas tendo duas colunas cada (proteína A, proteína B). Nessas tabelas foram armazenados os pares de proteínas com no mínimo 90% de identidade de sequência, identificados separadamente pelos três algoritmos. Depois, foi verificado quantos dos pares dos conjuntos identificados pelos programas heurísticos (GenPPI e BLASTp), também estavam no conjunto identificado pelo algoritmo exato Needleman-Wunsch. Para tanto, as tabelas desses dois programas foram comparadas de maneira independente por meio de *join* com a tabela do algoritmo Needleman-Wunsch. Assim, foi possível encontrar as intersecções dos conjuntos e verificar qual algoritmo heurístico se aproxima mais do exato Needleman-Wunsch em termos de exatidão na comparação par-a-par de proteínas.

4.1.3 Método de Validação da Hipótese de Correção Biológica Para Predições de PPI Feitas Pelo Programa Desenvolvido

O software proposto neste trabalho foi idealizado para construir redes de interação entre proteínas de genomas bacterianos. Entretanto, dados sobre duas características implementadas (vizinhança gênica conservada e perfil filogenético) são fundamentais em

estudos sobre filogenia de espécies bacterianas por evidenciarem o nível de conservação de espécies. Comparações de genes ou produtos gênicos como, por exemplo, as proteínas analisadas pelo nosso software, são pilares de uma análise de conservação de espécies. Portanto, uma das possibilidades de inferir sobre os níveis de correção e confiabilidade biológica das redes de PPI previstas pelo programa desenvolvido, é utilizar os dados que sustentam as previsões do mesmo, para comparar estatisticamente espécies evolutivamente correlacionadas. A expectativa dessa comparação é que genomas de espécies conhecidas como evolutivamente próximas apresentem medidas estatísticas muito similares e vice-versa. Com o intuito de verificar se os dados que a nossa ferramenta gera para prever redes de interação, são confiáveis em estudos de espécies de bactérias, 28 genomas do gênero *Dietzia*, 45 de *Rhodococcus*, 50 de *Corynebacterium* e 81 de *Aeromonas*, foram analisados pelo GenPPI. Os resultados obtidos pelo uso do software desenvolvido após aplicá-lo a esses genomas, são mostrados em três gráficos por gênero bacteriano, de modo a padronizar as análises. Esses gráficos apresentam medidas estatísticas utilizadas como medidas de avaliação para validar biologicamente as nossas previsões computacionais. Através dessas medidas estatísticas sobre os dados que sustentam as previsões do programa, consegue-se distinguir espécies e subespécies de gêneros bacterianos analisados. Até mesmo espécies incapazes de serem distinguidas pelos métodos bioquímicos, espécies do gênero bacteriano *Aeromonas*, por exemplo. Essa distinção bem sucedida através das análises estatísticas realizadas, é a evidência que valida a hipótese de correção e confiabilidade biológica para previsões do GenPPI. A explicação genérica sobre o que está sendo representado e analisado pelos três tipos de gráficos, é dada nas próximas três subseções. Esses gráficos foram feitos a partir dos dados de relatórios gerados pelo programa para os gêneros bacterianos analisados, utilizando a linguagem R com suporte de bibliotecas específicas para este fim.

4.1.3.1 Gráfico Boxplot Sobre o Perfil Filogenético das Proteínas Conservadas de Cada Genoma

Feito a partir de dados do relatório demonstrado na Subseção A.2.5, esse gráfico representa a quantidade de genomas do conjunto total analisado, que possuem as proteínas conservadas de cada genoma representado por um boxplot. Por esse motivo, o eixo Y (Genomes) está na escala de 0 a N, sendo N o total de genomas analisados. Nesse gráfico, uma mediana significa a quantidade de genomas em que as proteínas conservadas de um dado genoma, estão presentes. A largura de uma caixa de plotagem é proporcional à quantidade de proteínas conservadas de um genoma. A mediana da quantidade de genomas que possuem as proteínas conservadas de uma espécie, tende a ser uniforme em genomas de bactérias evolutivamente muito próximas ou da mesma espécie. Assim, podemos distinguir espécies de bactérias observando essa mediana dos genomas representados por esse gráfico.

4.1.3.2 Gráfico de Calor Sobre a Similaridade das Espécies de Bactérias Analisadas, com Base em seu Pan-Genoma

Feito a partir de dados do relatório demonstrado na Subseção A.2.2, esse gráfico demonstra a similaridade dos genomas das espécies de bactérias analisadas em uma execução do programa, com base no pan-genoma dessas espécies. É um gráfico de calor que estima o grau de similaridade dos genomas numa escala de 0 a 1. Um par de genomas com pouca ou nenhuma similaridade tem grau de similaridade próximo de 0 ou igual a 0, e um par de genomas muito similar tem grau de similaridade igual a 1 ou próximo de 1. A representação do grau de similaridade de um dado par de genomas, é feita por meio de um esquema de cores. Adotou-se a cor vermelha para significar pouca similaridade e branco significando genomas idênticos. A cor amarelo é um valor intermediário entre vermelho e branco. Assim, cores mais escuras e avermelhadas representam um menor grau de similaridade, e cores mais claras tendendo a amarelo e branco representam um maior grau de similaridade entre os genomas em questão. Além de demonstrar o grau de similaridade de genomas nesse gráfico de calor, também foi feito um agrupamento de genomas por grau de similaridade de modo a destacar grupos de genomas mais similares entre si. Tais grupos podem ser notados pela formação de quadrados perfeitos de cores claras (amarelo ou branco), ou através de um dendrograma incluído no gráfico para indicar quais genomas ou grupos de genomas são mais similares entre si.

4.1.3.3 Gráfico Boxplot da Quantidade de Genes Conservados na Vizinhança Gênica dos Genes Recorrentes, Considerando uma Janela de Expansão Fixa de Tamanho igual a 7

Feito a partir de dados do relatório demonstrado na Subseção A.2.3, esse gráfico é sobre a quantidade de genes conservados na vizinhança gênica dos genes recorrentes de cada genoma. Aqui temos uma janela de expansão fixa de tamanho igual a 7 genes vizinhos subsequentes dos genes recorrentes. São contabilizados quantos desses sete genes estão conservados na vizinhança dos genes recorrentes nos genomas onde estes estão presentes. Uma conservação gênica é registrada quando um dos sete genes vizinhos subsequentes de um gene recorrente, estiver conservado na vizinhança dele em dois genomas ou mais. Genomas muito similares podem possuir vizinhanças gênicas idênticas para genes recorrentes. Por esse motivo, quando há genomas muito similares no conjunto estudado, ao analisá-los com o GenPPI utilizando uma janela de expansão fixa de tamanho igual a sete, nesse gráfico os boxplots de genomas que possuem outro muito similar no conjunto de análise, podem apresentar mediana, o primeiro e terceiro quartil, bem como o valor máximo iguais a 7, exceto para alguns outliers. Isso significa que todos os 7 genes da janela de expansão estão conservados na vizinhança gênica dos genes recorrentes em pelo menos outro genoma do conjunto analisado. Assim, em vez de uma caixa de plotagem, é comum aparecer nesse gráfico apenas o indicativo da mediana, um traço vertical na escala

7 do eixo Y. Quanto mais extenso for um boxplot maior terá sido a quantidade de genes recorrentes com vizinhança gênica não conservada na extensão da janela de 7 genes, e vice versa. Por esse gráfico não há como saber quem são os genomas muito parecidos, mas apenas que há genomas muito parecidos, caso haja boxplots com mediana, o primeiro e terceiro quartil, bem como o valor máximo iguais a 7.

4.1.4 Método Para Comparação de Redes de Interação Geradas Pelo GenPPI Com Uma Rede da Ferramenta STRING

Para comparar redes de PPI geradas pelo nosso software, escolhemos o STRING (Ferramenta de Pesquisa para Recuperação de Genes em Interação) (SZKLARCZYK et al., 2019). O STRING (<https://string-db.org/>) é o principal software do estado da arte, para predição de interações entre proteínas. Visando comparar redes de interação geradas pelo GenPPI com uma rede do STRING, utilizou-se duas abordagens: comparação direta e indireta. Ao comparar diretamente redes de PPI, focamos em identificar interações de pares de proteínas em comum e específicas entre as redes comparadas. Já na comparação indireta o foco foi em características topológicas de redes complexas, que podem fornecer uma visão geral sobre a qualidade de uma rede de interação proteica, para realização de análises biológicas. As características topológicas avaliadas são definidas a partir de métricas como:

1. **Número de nós:** representa o número de proteínas interagentes identificadas;
2. **Número de arestas:** representa o número de interações proteicas preditas;
3. **Grau médio:** média aritmética do número de interações de todas as proteínas da rede. Indica o número de interações existentes em comparação com o número de proteínas. Baixos valores de grau médio simbolizam redes onde as interações entre proteínas se apresentam de uma forma mais distribuída proporcionando mais clareza para análises biológicas;
4. **Densidade:** razão entre o número total de interações proteicas e o número total de possíveis interações de acordo com a quantidade de proteínas de uma rede. Em relação a esta métrica, busca-se valores iguais ou menores aos encontrados na rede de referência do STRING. Isso porque altas densidades representam redes altamente conectadas, nas quais não se consegue inferir muitas informações. Valores de densidade menores que 0,1 são satisfatórios;

Acreditamos que estas quatro métricas podem fornecer uma visão geral da qualidade topológica de redes de interação proteína-proteína para realização de análises biológicas.

4.2 Experimentos

Nesta seção são apresentados os experimentos e análises de resultados que sustentam a hipótese de correção e confiabilidade biológica para as nossas previsões computacionais. Também experimentos para uma avaliação de resultados sobre a otimização de código realizada e a heurística proposta, além de uma comparação com a ferramenta STRING. Os experimentos descritos nesta seção foram realizados em um computador de configuração convencional: marca Samsung, modelo NP270E5K-KW2BR, com 8 GB de memória RAM e CPU dual core modelo Intel Core i5-5200U 64 bits, com clock de 2.20 GHz e Cache de 3 MB. As bases de dados (arquivos multi-fasta) com as sequências de proteínas dos genomas utilizados nos experimentos, foram obtidas através do site do NCBI.

4.2.1 Experimentos e Avaliação de Resultados Sobre a Otimização de Código Realizada

Nesta subseção é apresentado um experimento e avaliação de resultados sobre a otimização de código Lisp realizada para resolver problemas de estouro de memória. Nesse experimento utilizou-se uma base de dados constituída por 50 genomas de *Corynebacterium pseudotuberculosis*. Esses genomas foram processados por uma versão do programa sem otimização de código e por outra versão com otimização. Os resultados em termos de consumo de memória RAM, quantidade de ciclos executados pelo processador e o tempo de processamento das duas versões, estão apresentados na Tabela 7.

Tabela 7 – Resultados das versões do programa sem otimização de código e com otimização.

Métricas	Versão sem otimização	Versão com otimização
Alocação de memória RAM	436.65 mb	347.05 mb
Quantidade de ciclos de máquina	6.616.032.505.292	6.214.719.960.425
Tempo de processamento	00:50:23	00:47:19

Nesse experimento, a alocação de memória RAM realizada para processar a versão com otimização de código, representa uma queda razoável de 20,52% em relação à quantidade de memória alocada para processar a versão sem otimização de código. Com relação à quantidade de ciclos executados pelo processador e ao tempo de processamento, a versão com otimização apresenta uma queda sutil de 6,07% para ambas as métricas. Os resultados em termos de tempo de processamento sustentam uma das hipóteses levantadas neste trabalho. Isto é, a possibilidade de realizar previsões *ab initio* de redes de interação proteína-proteína em um tempo aceitável, não proibitivo. A prova é que mesmo para uma base de dados com uma quantidade expressiva de genomas, 50 genomas de *Corynebacterium pseudotuberculosis* contendo em média 2200 proteínas cada, o GenPPI realizou

suas predições com menos de uma hora de processamento. Tal agilidade é atribuída à heurística proposta.

4.2.2 Experimentos e Avaliação de Resultados Sobre a Heurística Proposta Para Comparação de Sequências Proteicas

Os experimentos realizados para avaliar a heurística proposta consistiram inicialmente em uma validação de resultados feita utilizando o algoritmo exato de comparação de sequências biológicas Needleman-Wunsch. E posteriormente, na comparação de desempenho com o BLASTp, o principal algoritmo heurístico do estado da arte para comparação de sequências de proteínas. As medidas de avaliação adotadas nessa comparação, foram: exatidão na comparação par-a-par de proteínas, tempo de processamento e uso de memória RAM.

4.2.2.1 Validação da Heurística Proposta

Inicialmente, o objetivo perseguido consistiu em encontrar valores para os parâmetros da heurística do GenPPI, que garantissem um percentual satisfatório de identidade mínima (90%) para todo par de proteínas classificado similar pela heurística. Para tanto, foi utilizada uma base de dados com 3988 sequências de aminoácidos de proteínas oriundas de um genoma de *Mycobacterium tuberculosis (H37Rv)*. Todo possível par de proteínas dessa base de dados, foi submetido à heurística do programa, para classificá-lo como um par similar ou não similar. As classificações positivas (classificações de pares similares), foram submetidas para análise do algoritmo exato de comparação de sequências biológicas Needleman-Wunsch. O objetivo foi estimar com o algoritmo exato, um percentual de identidade para todo par de proteínas classificado similar pela heurística. Por final, calculou-se algumas medidas estatísticas, como o percentual de identidade mínimo, máximo, médio e a mediana dos percentuais de identidade desses pares de proteínas. Assim, verificamos a qualidade dos resultados produzidos pelo uso de diferentes combinações de valores para os parâmetros da heurística (*d-limite* e *qtd-amin*).

A Tabela 8 apresenta algumas configurações de valores encontrados para os parâmetros da heurística, capazes de garantir uma exatidão satisfatória na classificação de pares de proteínas similares. Essa exatidão é refletida principalmente pelo percentual de identidade mínimo estimado pelo algoritmo Needleman-Wunsch, para os pares de proteínas classificadas como sendo similares pela heurística.

Como pode ser visto na Tabela 8, a variação de valores para os parâmetros da heurística (*d-limite* e *qtd-amin*), foi feita de maneira decrescente relaxando as configurações visando identificar conjuntos maiores de pares de proteínas similares. O percentual de identidade média estimado pelo algoritmo Needleman-Wunsch, para os conjuntos de pares similares identificados pela heurística, ficou acima de 99%. Qualquer das configurações de valores

Tabela 8 – Resultados obtidos com a combinação de diferentes valores para os parâmetros *d-limite* e *qtd-amin* da heurística do GenPPI.

Resultados por variação de valores para os parâmetros da heurística						
Legenda:		1. <i>d-limite</i> : limite tolerado na diferença de histogramas de aminoácidos para um par de proteínas. 2. <i>qtd-amin</i> : quantidade de aminoácidos com diferença de histogramas dentro do limite tolerado.				
Parâmetros da Heurística		Nº de pares similares classificados pela heurística	Estatísticas referentes às estimativas do algoritmo exato Needleman-Wunsh			
<i>d-limite</i>	<i>qtd-amin</i>		Identidade mínima	Identidade máxima	Identidade média	Mediana da identidade
0	26	336	100%	100%	100%	100%
0	25	336	100%	100%	100%	100%
0	24	360	97,96%	100%	99,95%	100%
0	23	366	96,94%	100%	99,91%	100%
0	22	368	96,94%	100%	99,9%	100%
0	21	369	94,68%	100%	99,87%	100%
0	20	370	91,75%	100%	99,83%	100%
1	26	360	97,87%	100%	99,95%	100%
1	25	370	92,55%	100%	99,84%	100%

que estão apresentadas nas colunas *d-limite* e *qtd-amin* da Tabela 8, garantiram essa média. O percentual de identidade mínima, que é o mais importante, ficou acima de 90% para todas as configurações de valores apresentados nessas colunas. Como configuração padrão do programa para os parâmetros da heurística, adotou-se os valores 1 e 25. Segundo a estimativa do algoritmo exato Needleman-Wunsch, estes garantem um percentual de identidade mínima igual a 92,55% para todo par de proteínas classificado similar pela heurística. Vale mencionar que o usuário pode configurar o GenPPI para fazer suas predições com qualquer uma dessas configurações de parâmetros. Também foram testadas outros valores para os parâmetros da heurística, mas os mesmos não estão inclusos na Tabela 8 por não terem gerado identidade mínima maior ou igual a 90%. Um exemplo ilustrativo da aplicação da heurística proposta, que demonstra com clareza o significado dos parâmetros *d-limite* e *qtd-amin*, foi apresentado na Tabela 3 da Seção 3.5.

4.2.2.2 Comparação da Heurística Proposta

Posteriormente, de acordo com o que está descrito na Subseção 4.1.2 referente ao método para avaliação da heurística proposta, foi feita uma comparação de desempenho da nossa heurística com o BLASTp. O objetivo foi verificar qual dos dois algoritmos heurísticos se aproxima mais do algoritmo exato para comparação de sequências biológicas Needleman-Wunsch. Para mensurar tal aproximação, a principal medida de avaliação adotada é a exatidão na comparação par-a-par de proteínas. Além da exatidão, também se compara o tempo de processamento e o consumo de memória RAM dos dois algoritmos heurísticos. Conforme está descrito na Subseção 4.1.2, vale lembrar que essas medidas de avaliação adotadas, foram calibradas com os resultados gerados pelos três algoritmos (Needleman-Wunsch, GenPPI e BLASTp), após aplicá-los a uma base de dados selecionada. A saber, 4026 sequências de aminoácidos de proteínas oriundas de um genoma de *Mycobacterium tuberculosis*. Dentre essas, identificou-se um conjunto de pares similares utilizando cada algoritmo. Assumimos como um par similar, duas proteínas com no mínimo 90% de identidade de sequência. Para estimar a exatidão dos dois algoritmos heurísticos (GenPPI e BLASTp), os resultados dos mesmos foram validados frente aos do algoritmo exato Needleman-Wunsch. Essa validação foi feita verificando quantos dos pares similares identificados por um algoritmo heurístico, encontravam-se no conjunto identificado pelo algoritmo exato. A Tabela 9 mostra os resultados obtidos pelos algoritmos heurísticos, frente aos do algoritmo exato.

Como pode ser visto na Tabela 9, fizemos várias execuções com o programa BLASTp variando seus parâmetros. O objetivo foi gerar resultados que se aproximassem o máximo possível dos obtidos pelo algoritmo exato Needleman-Wunsch, em termos de exatidão na comparação par-a-par de proteínas. Entre todas as execuções com o BLASTp, o melhor resultado identificou 590 pares como sendo similares, dentre os quais 522 foram validados pelo algoritmo exato Needleman-Wunsch. Ou seja, 68 pares julgados similares pelo

Tabela 9 – Resultados do GenPPI e BLASTp frente aos resultados do algoritmo exato Needleman-Wunsh.

Programa	Parâmetros	Tempo gasto	Memória (Mb)	Total de pares de proteínas identificados como similares (ident. mín. = 90%)	Total de pares validados pelo algoritmo Needleman-Wunsh
Needleman-Wunsh	-s BL50 -E 0.000001	00:01:23	153.4 mb	524	--
GenPPI	Default	00:00:05	153.4 mb	503	501
BLASTp	-matrix PAM250 -word_size 2 -evaluate 0.000001	00:09:19	99.71 mb	590	522
BLASTp	-matrix PAM250 -word_size 3 -evaluate 0.000001	00:05:35	107.38 mb	590	522
BLASTp	-matrix PAM250 -word_size 4 -evaluate 0.000001	00:14:19	337.48 mb	590	522
BLASTp	-matrix PAM250 -word_size 5 -evaluate 0.000001	Execução interrompida por estouro de memória.			
BLASTp	-matrix PAM70 -word_size 2 -evaluate 0.000001	00:03:28	99.71 mb	595	522
BLASTp	-matrix PAM70 -word_size 3 -evaluate 0.000001	00:01:08	99.71 mb	595	522
BLASTp	-matrix PAM70 -word_size 4 -evaluate 0.000001	00:02:46	153.4 mb	595	522
BLASTp	-matrix PAM70 -word_size 5 -evaluate 0.000001	00:39:12	1280.89 mb	595	522
BLASTp	-matrix PAM30 -word_size 2 -evaluate 0.000001	00:02:34	99.71 mb	597	522
BLASTp	-matrix PAM30 -word_size 3 -evaluate 0.000001	00:00:41	99.71 mb	597	522
BLASTp	-matrix PAM30 -word_size 4 -evaluate 0.000001	00:01:24	138.06 mb	597	522
BLASTp	-matrix PAM30 -word_size 5 -evaluate 0.000001	00:13:48	713.31 mb	597	522

BLASTp, não foram validados pelo algoritmo exato. Isso representa um erro de 11,53% nas comparações par-a-par feitas pelo algoritmo heurístico do BLASTp. Nesse sentido a nossa heurística apresentou o melhor resultado, pois dentre os seus 503 pares identificados como sendo similares, 501 foram validados pelo algoritmo exato Needleman-Wunsh. Ou seja, apenas 2 pares julgados similares pelo GenPPI, não foram validados pelo algoritmo exato. Isso representa um erro de apenas 0,39% vs 11,53% do BLASTp. É verdade que dentre os pares similares do conjunto encontrado pelo algoritmo exato (524 pares), o

BLASTp encontrou um pouco mais que o GenPPI, 522 vs 501 pares (uma aproximação de 99,62% vs 95,61% em relação ao total de pares similares encontrados pelo algoritmo exato). No entanto, o BLASTp gerou um número muito maior de pares identificados como sendo similares, não validados pelo algoritmo exato, 68 vs 2 da nossa heurística. Isso indica um maior percentual de erro para BLASTp e, conseqüentemente, menor exatidão para o mesmo. Assim, os resultados obtidos nesse experimento, evidenciam que o GenPPI produziu resultados mais satisfatórios frente ao algoritmo exato Needleman-Wunsh.

Também em termos de tempo de processamento, o software proposto neste trabalho demonstrou o melhor resultado. Ficou disparadamente na frente executando em apenas 5 segundos vs 41 do BLASTp considerando o seu melhor tempo entre todas as execuções realizadas nos testes. Apenas em termos de consumo de memória RAM, o GenPPI apresentou pior resultado em comparação com algumas execuções específicas do BLASTp. O custo computacional dos principais algoritmos do estado da arte para comparação de seqüências biológicas (BLASTp e Needleman-Wunsh), é refletido pelos seus tempos de processamento significativamente maiores, observados nesse experimento.

É importante salientar que a proposta deste trabalho referente a predição *ab initio* de PPI a partir de dados genômicos, somente se apresentou viável em termos de tempo de processamento, por causa da heurística proposta. Não sendo a mesma, certamente a solução seria optar por abordagens de aprendizado de máquina para prever interações de proteínas com algum percentual de acurácia satisfatório.

4.2.3 Experimentos e Avaliação de Resultados Sobre a Hipótese de Correção Biológica Para Predições de PPI Feitas Pelo Programa Desenvolvido

Nesta subseção é apresentada uma análise dos resultados obtidos após os experimentos realizados para validar a hipótese de correção e confiabilidade biológica para redes de PPI preditas pelo software desenvolvido. Tais experimentos consistiram na aplicação do GenPPI à genomas de gêneros bacterianos adotados como estudos de caso. Foram selecionados 28 genomas do gênero *Dietzia*, 45 de *Rhodococcus*, 50 de *Corynebacterium* e 81 de *Aeromonas* para serem submetidos às análises do nosso software. Portanto, são mostrados resultados obtidos particularmente para cada gênero de bactéria. Será mostrado que a ferramenta proposta, é capaz de fazer predições de interações proteicas biologicamente corretas e confiáveis. Considerando que esta é a hipótese central deste trabalho, o objetivo aqui é sustentar tal hipótese com base na construção de mapas de filogenia finamente detalhados para os genomas estudados. Será mostrado que a partir de dados gerados pelo GenPPI sobre o perfil filogenético das proteínas conservas e sobre o pan-genoma das espécies, podemos distinguir, por exemplo, entre biovars (subespécies) da espécie *Corynebacterium pseudotuberculosis* (BERNARDES et al., 2020). Além de possibilitarem a

separação ideal entre organismos procariotos de outros gêneros bacterianos.

Alguns resultados e conclusões desta subseção foram incluídos em um artigo que, até o mês de Dezembro de 2020, ainda estava sendo produzido pela rede RECOM composta por mais de 40 pesquisadores da área biológica fixados em diferentes regiões do Brasil. Esses resultados se mostraram importantes evidências para a RECOM, sobre o nível de conservação de genomas recentemente produzidos pela rede, ao compará-los com genomas previamente depositados por outros grupos de pesquisa em bancos de dados públicos. É importante salientar que nossas análises computacionais foram decisivas nas conclusões desse artigo sob redação, pois os métodos bioquímicos comumente utilizados para diferenciar bactérias, não possuem exatidão suficiente para dividir entre espécies do gênero bacteriano *Aeromonas*.

Os três estudos de caso a seguir foram realizados com o propósito de demonstrar que as redes de interação proteína-proteína obtidas através do uso do software proposto, possuem correção e confiabilidade biológica.

4.2.3.1 Estudo de Caso 1 – Gênero Bacteriano *Dietzia*

Dietzia tem causado várias infecções em humanos. Pesquisas fizeram novas classificações de algumas infecções que tinham sido causadas por *Rhodococcus*, porém mais tardiamente foi constatado que eram na verdade, infecções causadas por *Dietzia* (PILARES et al., 2010) (NIWA et al., 2012). Muitos trabalhos relatam que é preciso um teste mais eficaz para separar essas duas espécies, pois *Dietzia* ganhou uma grande importância em anos recentes, devido a quantidade de casos relatados em humanos. Na literatura é reportado com frequência, que esses dois gêneros de bactérias são comumente confundidos em laboratórios. A causa da confusão está na morfologia e aparência de colônias cultivadas de *Dietzia*, que são notavelmente semelhantes às de *Rhodococcus equi*.

A seguir é apresentada uma análise de resultados, embasada nos gráficos descritos na Subseção 4.1.3.

O eixo Y (Genomes) do gráfico da Figura 28, está na escala de 0 a 73, o total de genomas analisados. A mediana da quantidade de genomas que possuem as proteínas conservadas de genomas específicos (Genome), está uniforme ao longo da maioria das cepas de *Rhodococcus* (a mediana do número de genomas nos quais foram encontradas as proteínas conservadas de cada cepa de *Rhodococcus*, é de aproximadamente 35 genomas dentre os 73 analisados). O mesmo acontece com *Dietzia*, em que a mediana está próxima de 8 para a maior parte dos genomas. A mediana das duas espécies (*Dietzia* e *Rhodococcus*), no tocante à quantidade de genomas que possuem as proteínas conservadas dessas duas espécies, é muito diferente (35 versus 8). É verdade que *Dietzia* está representada com 28 genomas versus 45 de *Rhodococcus*. Independente da quantidade numérica, a proporção de conservação média das proteínas conservadas de *Dietzia*, nos genomas do conjunto total analisado, é menor que *Rhodococcus*. Podemos separar as duas

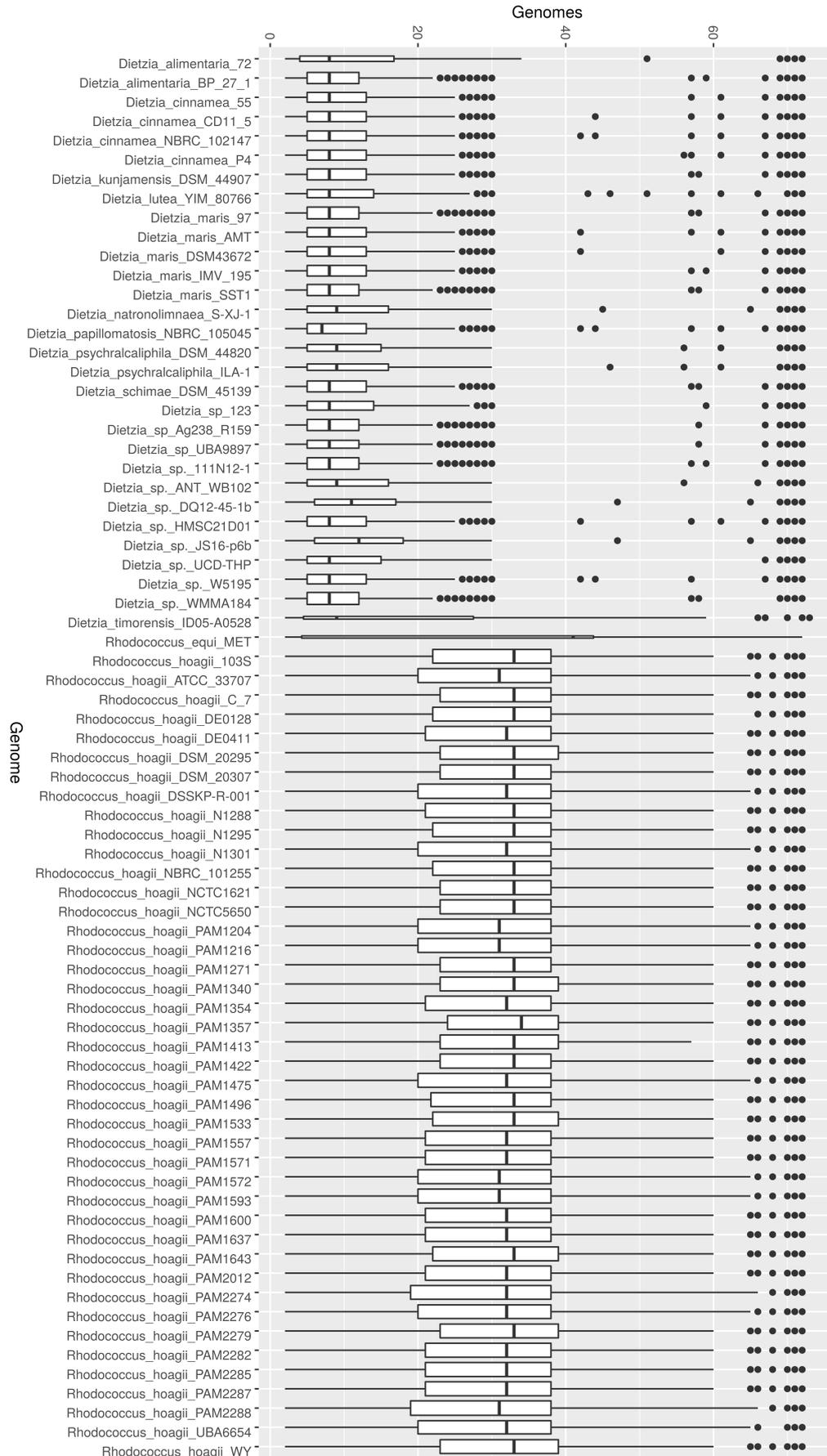


Figura 28 – Gráfico Boxplot Sobre o Perfil Filogenético das Proteínas Conservadas de Cada Genome (Genome)

espécies observando apenas a mediana representada nos boxplots dos genomas do gráfico da Figura 28. Apesar dessa diferença, existem outliers superiores representando que há genomas com proteínas conservadas em todos os outros. A importância dos resultados desse gráfico está na hipótese levantada pelo grupo RECOM, sobre a proximidade dos genomas de *Dietzia* com *Rhodococcus*. Esse gráfico sugere que, apesar de haverem proteínas conservadas em ambos os gêneros, os genomas desses dois gêneros são mais diferentes do que se pensou.

O grupo RECOM efetuou uma análise de 16S e do gene GyrB. Essa análise consiste em isolar genes presentes simultaneamente em todas as bactérias sob estudo, e mensurar o nível de separação desses genes, por exemplo, um experimento com gel de eletroforese. Por meio desse experimento de bancada, a RECOM concluiu ser possível separar as duas espécies, resultado que corrobora nossas análises computacionais de possibilidade de discernir entre cepas de *Dietzia* e de *Rhodococcus*.

O gráfico de calor da Figura 29, sobre a similaridade entre as espécies de bactérias do gênero de *Dietzia* e *Rhodococcus*, corrobora com os resultados do gráfico boxplot da Figura 28. Assim, conta-se com mais um indicativo de que a partir de dados gerados pelo GenPPI, agora no tocante ao pan-genoma das espécies analisadas, é possível distinguir entre espécies de *Dietzia* e *Rhodococcus*. Esse gráfico de calor compara genomas formando quadrados perfeitos de cor amarelo tendendo a branco para genomas mais similares entre si. Quanto mais próximo de branco é a cor mais similares são os genomas em comparação. Desse modo, é possível notar que os genomas de *Dietzia* são diferentes dos genomas de *Rhodococcus*, pois nesse gráfico de calor, não há cores claras nas posições de comparação (linha e coluna) dos genomas desses dois gêneros bacterianos (lateral inferior direita do gráfico). Tal diferenciação também pode ser percebida observando o dendrograma de similaridade entre os genomas analisados. Os ramos iniciais do dendrograma, indicam duas classes gerais de genomas. Essas representam, exatamente, os agrupamentos dos genomas de *Dietzia* e de *Rhodococcus*. Diferente dos genomas do gênero de *Rhodococcus*, que formaram um quadrado perfeito de cor amarelo representando um grau de similaridade alto entre os mesmos, isso não ocorre para os genomas de *Dietzia* (lateral inferior esquerda do gráfico). Esse fato representa uma menor similaridade entre os genomas desse último gênero. Também observou-se a formação de quadrados independentes de cores mais claras entre os genomas de *Dietzia*. Tal evidência indica a possibilidade de subespécies entre os genomas de *Dietzia*. Outro indicativo nesse sentido é o dendrograma que possui ramos menos homogêneos para os genomas de *Dietzia*.

Como dito na Subseção 4.1.3.3, o foco do gráfico da Figura 30 é mostrar se há genomas muito similares entre si dentre os genomas incluídos em uma análise. A evidência disso é cada boxplot ser pouco extenso, pois quanto mais extenso for o boxplot de um genoma, maior terá sido sua quantidade de genes recorrentes com vizinhança gênica não conservada na extensão da janela fixa de sete genes. Isso representa um baixo nível de similaridade

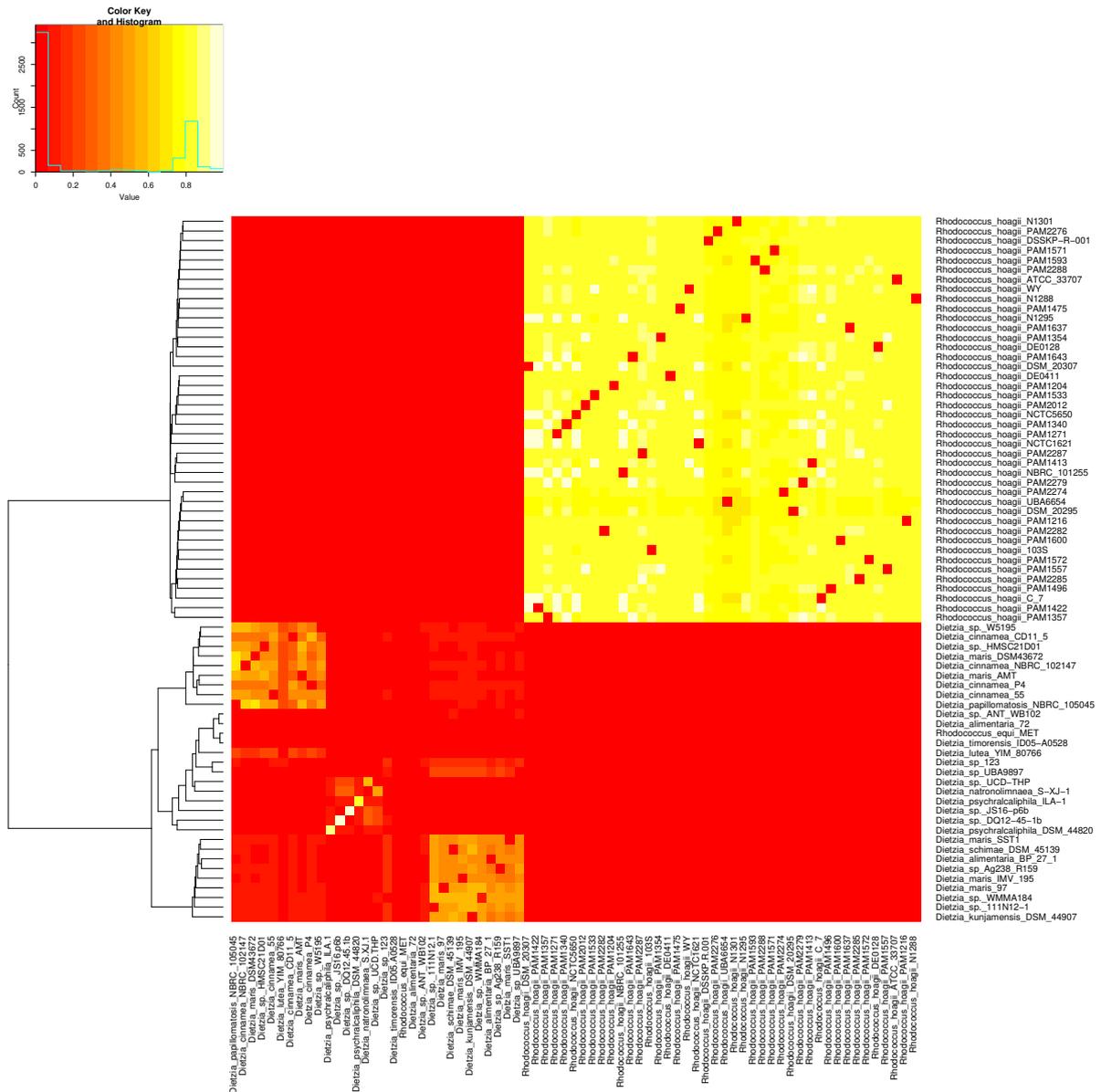


Figura 29 – Gráfico de calor sobre a similaridade das espécies de bactérias analisadas, com base em seu pan-genoma

entre os genomas. Por esse gráfico não há como saber quais são os genomas muito similares, mas apenas que há genomas muito similares entre os que foram analisados pelo programa. No gráfico da Figura 30, a maior parte dos genomas de *Rhodococcus* apresentam mediana, o primeiro e terceiro quartil, bem como o valor máximo iguais a 7. Por esse motivo, em vez de uma caixa de plotagem, esses genomas possuem apenas o indicativo da mediana, um traço vertical na escala 7 do eixo Y. Ou seja, eles são muito parecidos com outro ou outros genomas de *Rhodococcus*. Apenas cinco genomas de *Rhodococcus* contrariaram esse comportamento dentre os genomas dessa espécie. Em contrapartida,

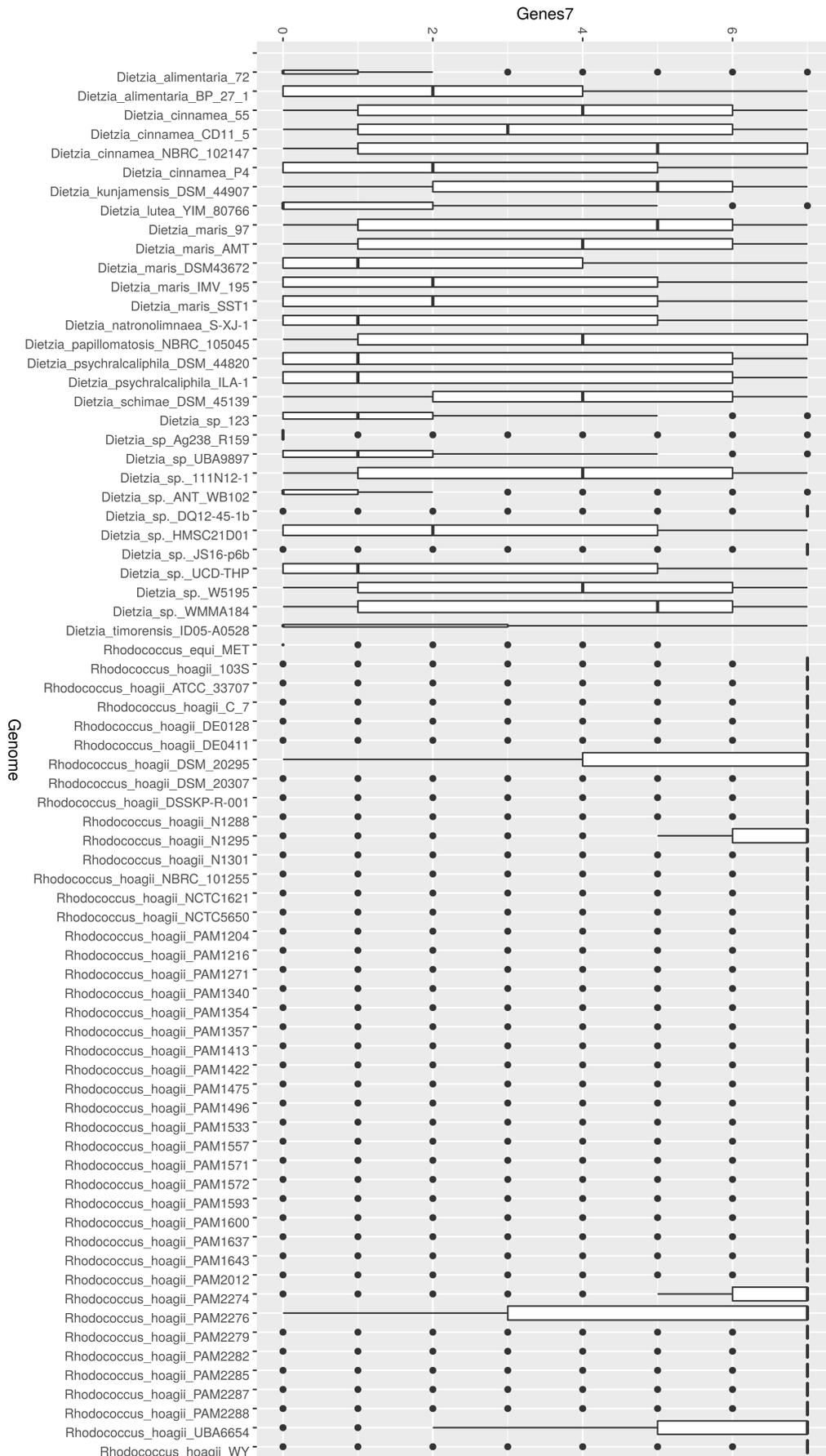


Figura 30 – Gráfico boxplot da quantidade de genes conservados na vizinhança gênica dos genes recorrentes, considerando uma janela de expansão fixa de tamanho igual a 7

apenas dois genomas de *Dietzia* apresentam esse comportamento, os genomas com os nomes *Dietzia*_sp._DQ12-45-1b e *Dietzia*_sp._JS16-p6b. Assim, foi possível indicar com facilidade que esses dois são os genomas evolutivamente mais próximos dentre os de *Dietzia* do conjunto analisado. Os demais genomas de *Dietzia* estão bastante diferentes entre si, com limites variando de zero a sete. A alta variação dos limites inferiores e superiores das caixas de plotagem desses genomas no gráfico da Figura 30, é outro indicativo que os genomas de *Rhodococcus*, são mais "homogêneos" do que os genomas de *Dietzia*. Esse resultado corrobora com a análise do gráfico de calor da Figura 29 onde os ramos do dendrograma referentes aos genomas de *Rhodococcus*, são mais homogêneos entre si do que os ramos dos genomas de *Dietzia*. Em biologia, o contrário de homogêneo é algo "plástico". Dessa forma, o GenPPI está evidenciando uma maior plasticidade genômica para o gênero *Dietzia* de bactérias, quando comparado com o gênero *Rhodococcus*.

4.2.3.2 Estudo de Caso 2 – Gênero Bacteriano *Corynebacterium*

Dentre as bactérias desse gênero a mais conhecida do grande público é a *Corynebacterium diphtheriae* por causar a doença da difteria em humanos. Inclusive, houve um surto dessa doença na Rússia durante a década de 1990, por conta da expiração da proteção da vacina em pessoas acima dos 50 anos de idade. Apesar desse ilustre representante do gênero, a espécie *C. pseudotuberculosis* também é conhecida no meio agropecuário em praticamente todos os países com clima predominantemente quente. No Brasil ela causa uma doença conhecida como Linfadenite Caseosa que acomete cabras e ovelhas causando perdas econômicas para pequenos criadores do norte e nordeste do Brasil, por tornar imprópria para consumo, a carne e a pele dos animais. As espécies de *C. pseudotuberculosis* e *C. diphtheriae* possuem grande semelhança tanto no modo de infecção quanto em seus conteúdos gênicos (GUARALDI; HIRATA; AZEVEDO, 2014), motivo pelo qual foram selecionadas para análise nesse estudo de caso.

Na sequência é apresentada uma análise de resultados baseada nos gráficos descritos na Subseção 4.1.3. Para tanto, 50 genomas de *Corynebacterium* foram computados (7 da espécie *C. Diphtheriae*, 16 de *C. pseudotuberculosis* e 27 de outras espécies de *Corynebacterium*).

O gráfico boxplot da Figura 31, é muito esclarecedor para quem é um estudioso do grupo *Corynebacterium*. O motivo é que o mesmo destaca fatos conhecidos para esses estudiosos, utilizando tão somente, dados sobre os perfis filogenéticos das proteínas conservadas entre as espécies e cepas. Nesse gráfico, conforme descrito na Subseção 4.1.3.1, uma mediana significa a quantidade de genomas que possuem as proteínas conservadas de um genoma específico. A largura de uma caixa de plotagem é proporcional à quantidade de proteínas conservadas de um genoma específico. As nossas análises computacionais pelo método de perfil filogenético conservado, obtiveram o discernimento entre as linhagens do biovar *ovis* e *equi* de *C. Pseudotuberculosis* (BERNARDES et al., 2020). Um

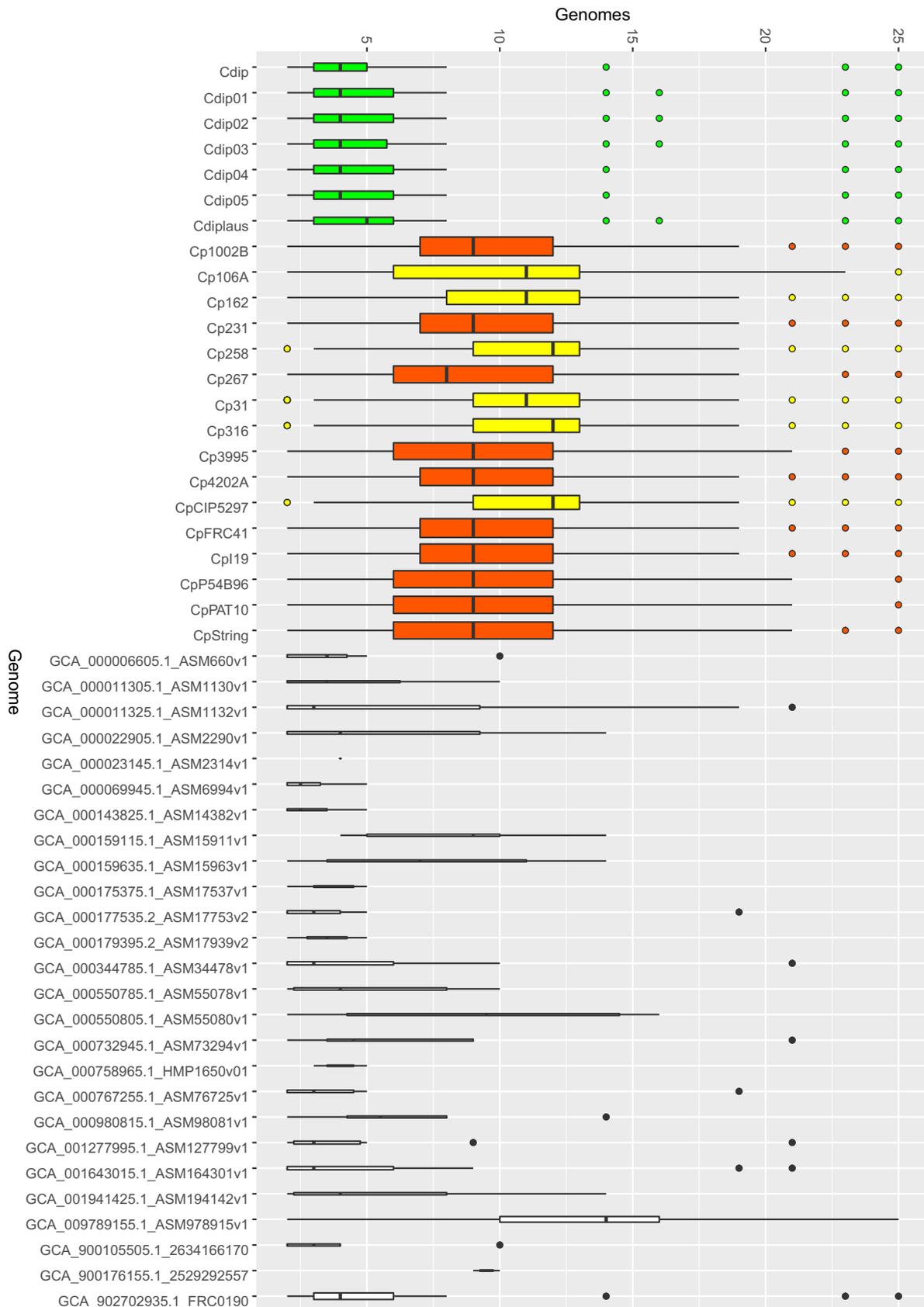


Figura 31 – Gráfico Boxplot Sobre o Perfil Filogenético das Proteínas Conservadas de Cada Genoma (Genome)

biovar é uma variante procariótica que difere fisiologicamente ou bioquimicamente de outras linhagens de uma espécie específica. No gráfico da Figura 31, essa diferenciação foi evidenciada pela altura da mediana e dos primeiros e terceiros quartis das caixas de plotagem desses genomas. O biovar *equi* (caixas de plotagem com a cor amarelo) fica com o primeiro quartil das caixas de plotagem de seus genomas, alinhado próximo à mediana das caixas de plotagem dos genomas pertencentes ao biovar *ovis* (caixas de plotagem com a cor laranja). Os genomas do biovar *equi* que se enquadram nesse cenário, são os seguintes: Cp162 (oriundo de um camelo no Egito), Cp258 (oriundo de um cavalo nos EUA), Cp31 (oriundo de um búfalo), Cp316 (oriundo de um cavalo nos EUA) e CpCIP5297 (oriundo de um cavalo no Quênia). A exceção à essa regra foi o genoma Cp106A (oriundo de um cavalo nos EUA), que apresentou o primeiro quartil mais próximo das linhagens do biovar *ovis*. Entretanto, a mediana do genoma Cp106A está mais próxima da mediana dos genomas do biovar *equi*. As caixas de plotagem mais largas são dezesseis ao todo (caixas de cores amarelo e laranja), e são todas da espécie *C. pseudotuberculosis*. As caixas de plotagem de genomas da espécie *C. diphtheriae* (caixas de plotagem com a cor verde), são sete e possuem largura menor do que as de *C. pseudotuberculosis*. Podemos discernir essas duas espécies observando apenas a mediana que está aproximadamente igual a 4 para os genomas de *C. diphtheriae* e maior que 8 para os de *C. pseudotuberculosis*. As demais espécies, inclusive por estarem menos representadas nesse conjunto de genomas, não mostraram conservação proteica expressiva e, portanto, possuem caixas de plotagem com pequena largura. Essas demais espécies possuem apenas um genoma representando-as nesse conjunto de 50 genomas do gênero *Corynebacterium*. Essa baixa representatividade faz com que a quantidade de proteínas conservadas dessas espécies, seja apenas o comum ao gênero *Corynebacterium*. Diferente das espécies *C. diphtheriae* e *C. pseudotuberculosis* que, por estarem numericamente mais representadas, mostraram um número maior de proteínas conservadas entre os genomas sob análise. A espécie *C. diphtheriae* fora inclusive utilizada como referência para ajudar a montar o primeiro desses 16 genomas de *C. pseudotuberculosis*, a linhagem Cp1002B. À época acreditava-se que as duas espécies eram muito similares. Ao final da primeira montagem, concluiu-se que essas espécies possuem um nível de similaridade que fica acima de 60% ao nível proteico.

No gráfico de calor da Figura 32 mais uma vez os resultados das análises computacionais realizadas, possibilitaram a diferenciação de espécies de *Corynebacterium*, agora com base no pan-genoma dessas espécies. Nesse gráfico percebe-se claramente dois grupos distintos referentes aos genomas de *C. Pseudotuberculosis* e *C. Diphtheriae*. A diferenciação dos grupos de genomas dessas duas espécies, pode ser feita observando a formação de dois quadrados de cor amarelo nesse gráfico. O quadrado de maior tamanho localizado na parte superior direita do gráfico é referente aos genomas de *C. Pseudotuberculosis*, e o quadrado amarelo de menor tamanho é referente aos genomas *C. Diphtheriae*. Assim, evidencia-se a distinção entre os genomas dessas duas espécies no gráfico de calor da Figura 32. Além

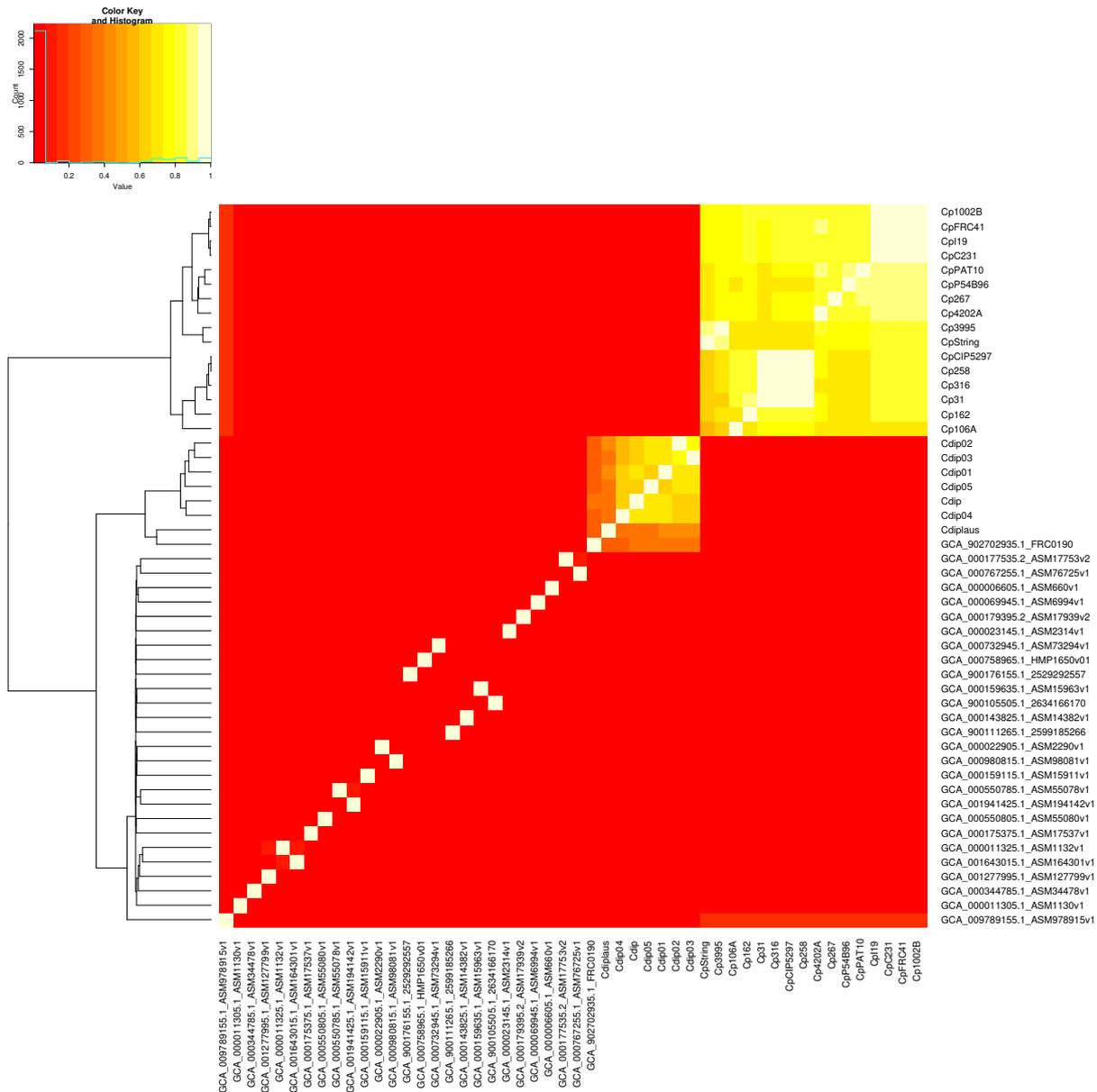


Figura 32 – Gráfico de calor sobre a similaridade das espécies de bactérias analisadas, com base em seu pan-genoma

da diferenciação das duas espécies, esse gráfico de calor possibilita também a separação entre os genomas do biovar *ovis* e *equi* de *C. Pseudotuberculosis*. Tal discernimento pode ser obtido observando o dendrograma do agrupamento por similaridade de genomas. Os genomas do biovar *ovis* e *equi*, são agrupados em dois grupos distintos imediatos à ramificação inicial direita. Os genomas do biovar *ovis* são: Cp1002B, CpFRC41, CP119, CpC231, CpPAT10, CpP5B96, Cp267, Cp4202A, Cp3995 e CpString. Estes encontram-se nesta mesma ordem nas linhas do gráfico de calor da Figura 32, começando de cima para baixo. Na sequência, os genomas do biovar *equi* seguem nessa ordem: CpCIP5297, Cp258, Cp316, Cp31, Cp162, e Cp106A. Ainda sobre o dendrograma, outro detalhe inte-

ressante pode ser notado nas ramificações desses dois grupos de genomas, biovar *ovis* e *equi*. Percebe-se a formação de subgrupos entre os genomas de cada um desses biovars. Essa formação de subgrupos entre os genomas do biovar *ovis* e *equi*, também é evidenciada pela formação de quadrados brancos dentro do quadrado amarelo de maior tamanho. Tais evidências indicam uma possível existência de subespécies para os genomas do biovar *ovis* e *equi*. Entretanto, como a definição de subespécies carece de outras análises, esperamos uma análise mais aprofundada para levantar essa hipótese.

Alguns casos são dignos de nota no gráfico de calor da Figura 32, por exemplo: (i) O genoma identificado como GCA_902702935.1_FRC0190 é referente à espécie *C. rouxii* (*high GC Gram+*). Esse genoma apresentou alta similaridade ao nível proteico com o agrupamento de *C. diphtheriae*. Apenas analisando os dados desse gráfico seria provável que esta espécie fosse de fato uma *C. diphtheriae*, em vez da espécie *C. rouxii*. Além dessas análises, recentemente a literatura especializada nesses organismos definiu a espécie *C. rouxi* como sendo um novo membro da espécie *C. diphtheriae* (BADELL et al., 2020). (ii) O genoma identificado como GCA009789155.1_ASM978915v1 é referente à espécie *Corynebacterium ulcerans strain MRI49*. De acordo com o gráfico da Figura 32, esse genoma apresentou similaridade ao nível proteico com os genomas de *C. pseudotuberculosis*. Assim, uma vez que se percebe uma leve coloração laranja no gráfico de calor, nos pontos de comparação desse genoma com os genomas de *C. pseudotuberculosis*, acredita-se que esta seja de fato uma espécie com alguma similaridade proteica para com *C. pseudotuberculosis*. Inclusive, a literatura descreve as espécies *C. pseudotuberculosis* e *C. ulcerans* como sendo evolutivamente relacionadas (BUSCH et al., 2019) (CLAVERYS et al., 1995).

No gráfico da Figura 33, a quantidade de genes conservados em uma janela de sete genes vizinhos subsequentes dos genes recorrentes, evidenciou uma alta similaridade entre os 16 genomas dos biovars *ovis* e *equi* de *C. pseudotuberculosis* (caixas de plotagem em cores laranja e amarelo, respectivamente). Excetuando-se o genoma Cp106A do biovar *equi* e o genoma CpString do biovar *ovis*, a mediana dos outros 14 genomas de *C. pseudotuberculosis* permaneceu no nível máximo da janela explorada, uma janela de sete genes. Em suma, dos 16 genomas de *C. pseudotuberculosis*, apenas dois ficaram com a mediana abaixo de sete. O motivo por quase todos genomas de ambos os biovars *ovis* e *equi*, estarem com uma mediana no valor máximo do eixo do Y, não possibilitando a distinção desses dois biovars, é decorrente da quantidade de sete que é relativamente pequena para uma janela fixa de contabilização de vizinhança gênica conservada.

Para tentar distinguir entres os biovars *ovis* e *equi* de *C. pseudotuberculosis*, pela característica de vizinhança gênica conservada, foram feitos testes de expansão dinâmica para os 50 genomas de *Corynebacterium* analisados nesta subseção. Esses testes geraram o gráfico da Figura 34 exclusivo para esse estudo de caso.

Quando o programa é executado sem restrição de uma janela máxima para análise de vizinhança gênica conservada, com aumentos progressivos até que a qualidade de conser-

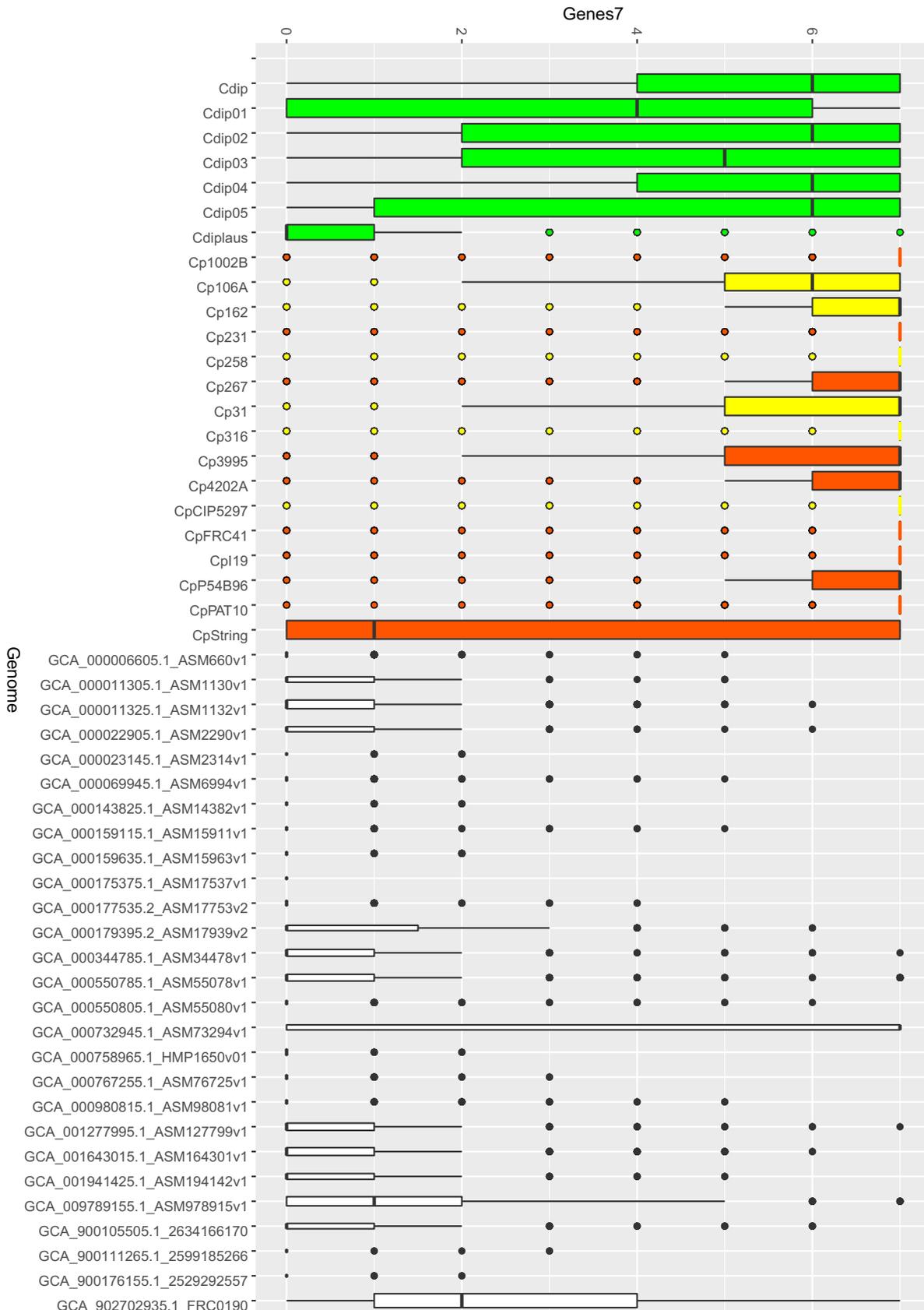


Figura 33 – Gráfico boxplot da quantidade de genes conservados na vizinhança gênica dos genes recorrentes, considerando uma janela de expansão fixa de tamanho igual a 7

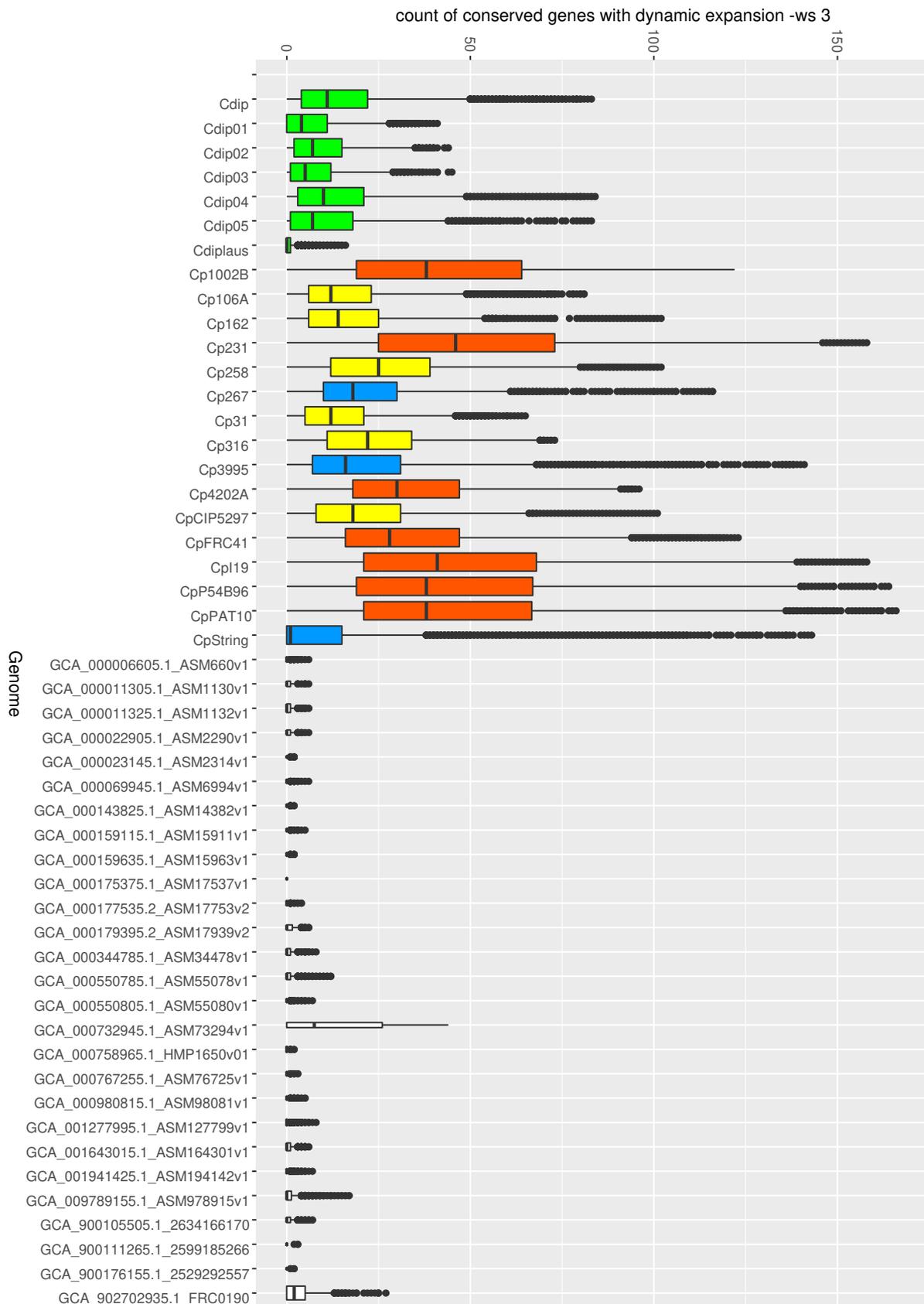


Figura 34 – Gráfico boxplot da quantidade de genes conservados na vizinhança gênica dos genes recorrentes, considerando uma janela dinâmica com intervalo igual a 3.

vação diminua, chamamos esse processo de expansão dinâmica (parâmetros: -expt 'dynamic' e -ws que significa um intervalo tolerado de genes subsequentes não conservados, para que a expansão continue). Com um intervalo igual a 3 (-ws 3), o gráfico da Figura 34 demonstra que a mediana do número de genes conservados na vizinhança dos genes recorrentes, mostrou uma alta similaridade entre os genomas do biovar *equi* da espécie *C. pseudotuberculosis*, quais sejam: Cp106A, Cp162, Cp258, Cp31, Cp316 e CpCIP5297 (caixas de plotagem com a cor amarelo). A mediana dos seis genomas equinos (biovar *equi*) permaneceu abaixo de 25 genes conservados. No gráfico da Figura 34, assim como os seis genomas do biovar *equi*, três do biovar *ovis* possuem boxplots com a mediana abaixo de 25, os genomas Cp267, Cp3995 e CpString (caixas de plotagem com a cor azul). As caixas de plotagem de cores azul e laranja são todas do biovar *ovis*. Quando o GenPPI foi executado com um intervalo de expansão dinâmica igual a 1 (-ws 1), observou-se sete entre dez genomas de *C. pseudotuberculosis* do biovar *ovis* cujas medianas se aproximaram dos valores dos genomas do biovar *equi*, não possibilitando assim, a separação dos dois biovares (dados não exibidos). Aumentando o intervalo da expansão dinâmica, por exemplo, para -ws 5 e -ws 7, não houve alterações em relação ao resultado com -ws 3 (dados não exibidos). Assim, o valor que melhor criou a separação dos biovares *ovis* e *equi*, no tocante à conservação da vizinhança dos genes recorrentes, foi um intervalo de expansão dinâmica igual a 3, valor derivado da experimentação e comparação entre resultados.

Ainda sobre o gráfico da Figura 34, utilizando a expansão dinâmica, os sete genomas de *C. diphtheriae* (caixas de plotagem de cor verde) apresentam medianas menores que as menores obtidas para a maioria das espécies de *C. pseudotuberculosis*. A mediana das espécies de *C. diphtheriae* manteve-se inferior à mediana da maioria das espécies de *C. pseudotuberculosis*. O genoma Cdiplaus, o sétimo contando de cima para baixo nesse gráfico, está com uma mediana bem abaixo dos outros genomas de *C. diphtheriae*. Isso indica que esse genoma tido como uma espécie de *C. diphtheriae*, possa ser na verdade de outra espécie. Tal indicação foi validada pela literatura recente que reportou o genoma Cdiplaus como sendo uma espécie de *Corynebacterium belfantii* ao invés de *C. diphtheriae* (BADELL et al., 2020). Assim, há evidências para acreditar que essa análise apresenta uma classificação correta para todos os genomas da espécie *C. diphtheriae*.

Nos gráficos das Figuras 33 e 34, são encontradas medianas com valores próximos a zero para genomas que possuem apenas um representante por espécie no conjunto analisado.

Nesta subseção a análise de resultados do gráfico de caixas da Figura 31, mostrou que os dados gerados pelo GenPPI para inferir interações de proteínas por perfil filogenético conservado, estão de acordo com o conhecimento literário de espécies e biovares (BERNARDES et al., 2020). A análise de resultados do gráfico de calor da Figura 32, mostrou a garantia de geração de um pan-genoma com capacidade de criar agrupamentos de filogenia coerentes com o conhecimento biológico de espécies bacterianas. Inclusive, evidenciou descobertas recentes, a recomendação de mudança de nomenclatura da espécie

C. rouxi para *C. diphtheriae* (BADELL et al., 2020). A análise do gráfico de caixas da Figura 34, mostrou que o software desenvolvido consegue representar vizinhanças gênicas conservadas, em conformidade com o conhecimento literário recente das espécies bacterianas analisadas. Um exemplo disso é o genoma Cdiplaus sempre tido como uma espécie de *C. diphtheriae*, mas que segundo análises feitas a partir de dados gerados pelo programa desenvolvido, seria na verdade um genoma de outra espécie, conforme corrobora (BADELL et al., 2020). Dessa forma valida-se a hipótese principal deste trabalho, ou seja, a possibilidade de correção e confiabilidade biológica para predições feitas pelo GenPPI. Nossas predições computacionais representaram com um alto grau de confiança, relações evolutivas de espécies e subespécies de bactérias sob análise. Portanto, acredita-se que redes de interação proteína-proteína geradas pelo software de bioinformática proposto, podem apresentar algum significado biológico.

4.2.3.3 Estudo de Caso 3 – Gênero Bacteriano *Aeromonas*

Recentemente, o grupo RECOM sequenciou e montou 69 genomas do gênero *Aeromonas*. Esse gênero causa doenças em peixes e, conseqüentemente, perdas econômicas para apicultores situados na bacia do rio São Francisco na região nordeste do Brasil. Considerando que a apicultura é uma importante atividade econômica para essa região do nosso país, é de grande importância caracterizar devidamente os genomas desse gênero, bem como encontrar alvos vacinais para o desenvolvimento de métodos de prevenção de doenças na apicultura. A proposta deste trabalho serve a ambos os propósitos, visto que a base da construção de redes de PPI, passa por levantar características que também permitem classificar espécies. Assim sendo, nossas análises computacionais ajudaram o grupo RECOM a classificar espécies de bactérias inéditas do gênero *Aeromonas*. No mês de Julho de 2020, esse grupo de pesquisa começou a utilizar redes de interação geradas no âmbito deste trabalho, com o intuito de procurar por alvos vacinais para essas novas espécies de bactérias. No entanto, até o mês de Junho de 2021 não havia resultados concretos sobre essa pesquisa.

Nesses gráficos, os genomas sequenciados pela RECOM se diferenciam dos demais que foram sequenciados por outros grupos, por meio dos prefixos dos nomes. Os nomes dos genomas sequenciados por outros grupos, começam com o prefixo GCA. Todos os demais foram sequenciados e montados pela RECOM. Há quatro espécies distintas de *Aeromonas* publicadas pela RECOM: *A. hydrophila*, *A. jandaei*, *A. dhakensis* e *A. veronii*.

Os resultados gerados em decorrência das análises computacionais realizadas para o conjunto de genomas de *Aeromonas*, são apresentados na sequência de acordo com os gráficos descritos na Subseção 4.1.3.

No gráfico da Figura 35 é possível distinguir com clareza as espécies *A. dhakensis* e *A. hydrophila* observando apenas a mediana representada nos boxplots dos genomas desse gráfico. Entretanto, para as espécies *A. jandaei* e *A. veronii*, essa distinção não

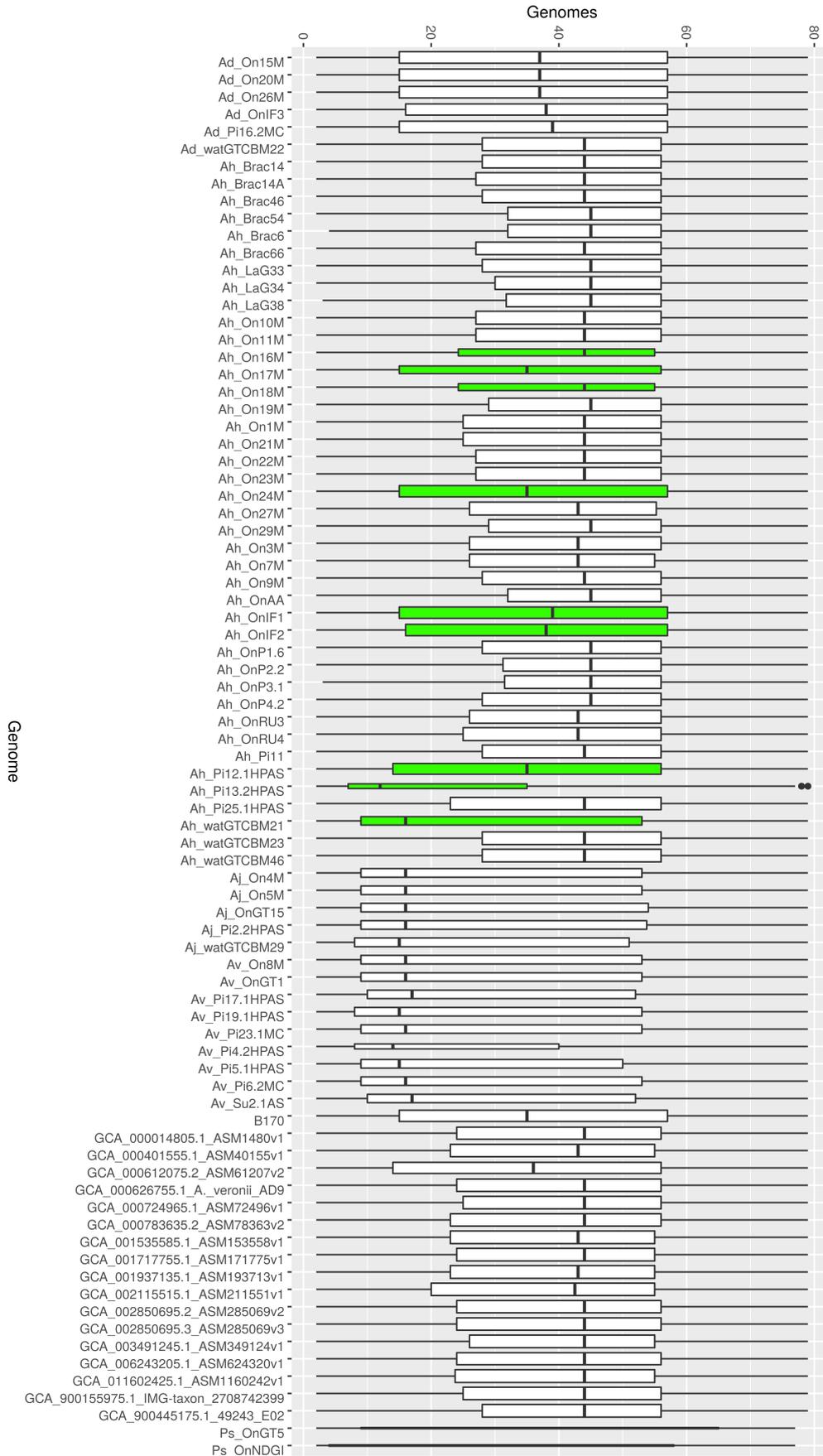


Figura 35 – Gráfico Boxplot Sobre o Perfil Filogenético das Proteínas Conservadas de Cada Genoma (Genome)

é perceptível através desse gráfico. Há muita dificuldade para separar essas espécies em testes bioquímicos. Não está descartada a possibilidade de que os testes bioquímicos realizados pelo grupo RECOM, tenham caracterizado de forma equivocada alguns dos genomas ora denominados *A. jandaei* ou *A. veronii*. Até o mês de Junho de 2020, diversas outras abordagens computacionais estavam sendo avaliadas para tentar se chegar a um consenso sobre esse dilema das espécies *A. jandaei* e *A. veronii*. Também sobre 9 genomas pontuais (boxplots de cor verde) em meio aos genomas de *A. hydrophila*, que destoaram do perfil do restante, quais sejam: Ah_On 16M, Ah_On17M, Ah_On18M, Ah_On24M, Ah_On1F1, Ah_On1F2, Ah_Pi12.1HPAS, Ah_Pi13.2HPAS e Ah_watGTCBM21.

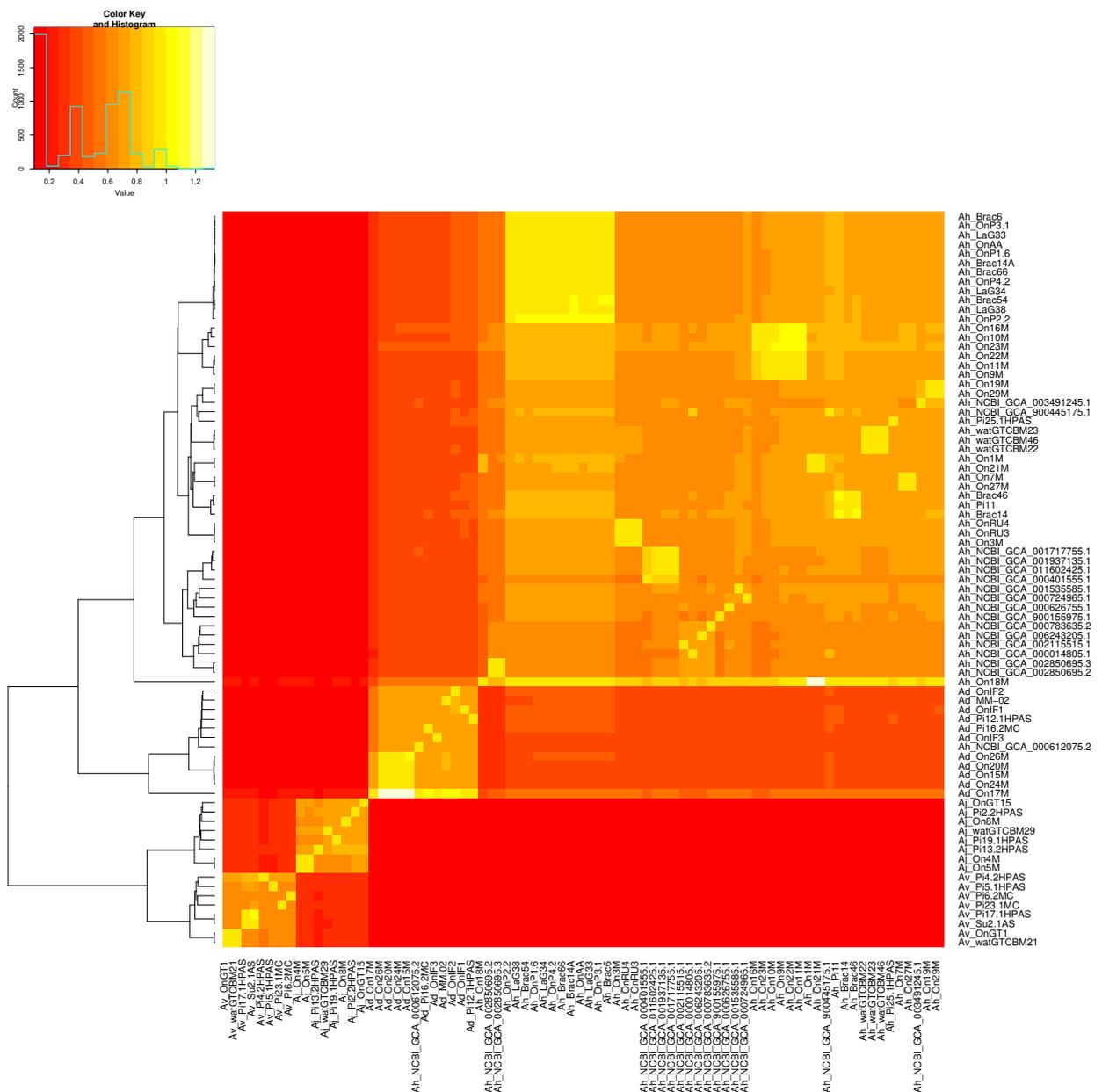


Figura 36 – Gráfico de calor sobre a similaridade das espécies de bactérias analisadas, com base em seu pan-genoma

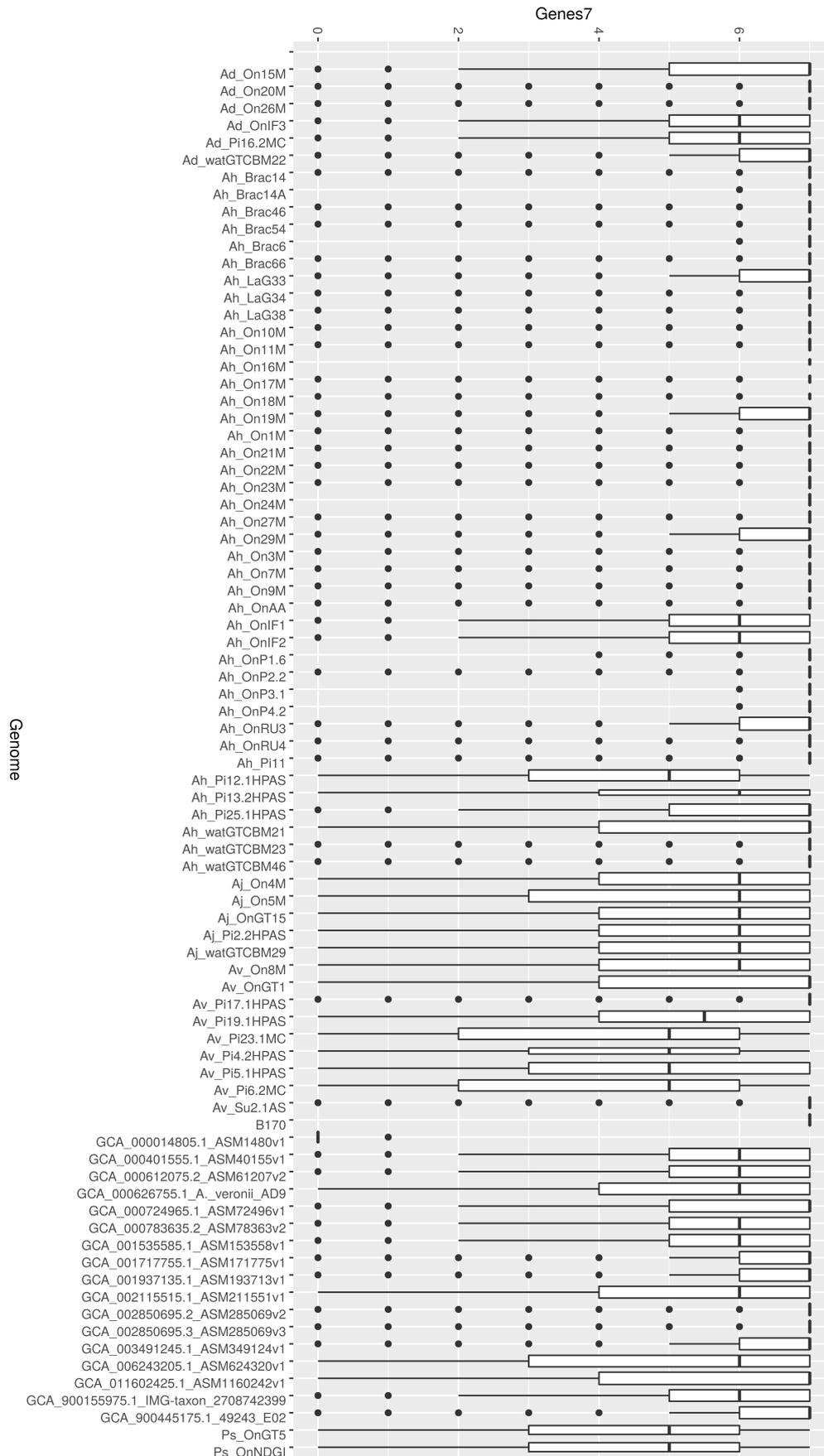


Figura 37 – Gráfico boxplot da quantidade de genes conservados na vizinhança gênica dos genes recorrentes, considerando uma janela de expansão fixa de tamanho igual a 7

Diferente do gráfico da Figura 35, no qual não foi possível diferenciar as espécies de *A. jandaei* e *A. veronii*, tal diferenciação foi possível através da Figura 36 referente ao gráfico de calor sobre a similaridade das espécies baseada em seu pan-genoma. Nesse gráfico pode-se observar que os genomas dessas duas espécies, estão distribuídos em ramos distintos do dendrograma (extremidade inferior esquerda do gráfico). Isso evidencia a diferença existente entre os genomas de *A. jandaei* e *A. veronii*. Outra forma de fazer distinção entre espécies nesse gráfico, é observar a formação de quadrados com cores claras nas posições de comparação (linha e coluna) dos genomas. Nesse sentido, também há evidência de distinção entre as espécies de *A. jandaei* e *A. veronii*, pois pode-se observar na parte inferior esquerda desse gráfico de calor, a formação de quadrados de cor amarelo para os genomas de cada uma dessas duas espécies. Observando tais características no gráfico da Figura 36, também é possível distinguir entre as outras espécies de *Aeromonas* analisadas.

No gráfico da Figura 37, a maior parte dos genomas de *A. hydrophila* (genomas com prefixo Ah), apresentaram mediana, o primeiro e terceiro quartil, bem como o valor máximo iguais a 7 em seus boxplots. Isso significa que o nível de conservação gênica da vizinhança de até 7 genes subsequentes aos genes recorrentes, é surpreendentemente alto para a maioria dos genomas de *A. hydrophila* em relação a pelo menos algum outro genoma desse conjunto analisado. Uma plasticidade (heterogeneidade) maior é observada nos genomas de *A. jandaei* (prefixo Aj) e *A. veronii* (prefixo Av).

Nesta seção foram apresentados os resultados obtidos após computar genomas dos três gêneros bacterianos selecionados como estudos de caso, a saber, *Dietzia*, *Corynebacterium* e *Aeromonas*. Na próxima seção, compara-se redes de interação geradas neste trabalho, com uma rede do principal software de predição de PPI do estado da arte, o STRING.

4.2.4 Comparação de Redes de Interação Geradas Pelo GenPPI Com Uma Rede do STRING

O software autônomo desenvolvido conta com uma grande variedade de parâmetros para gerar uma Rede de Interação (RI). Para comparar redes geradas pelo GenPPI, selecionamos uma rede fornecida pela ferramenta STRING (<https://string-db.org/>), para o genoma de *Corynebacterium pseudotuberculosis strain:Ft_2193/67*. Visando criar redes dessa mesma cepa com o nosso software, 50 genomas de *Corynebacterium* foram submetidos a execuções do GenPPI com diferentes configurações de parâmetros. Tais parâmetros foram escolhidos de modo a dividir as análises entre os dois métodos implementados para inferir PPIs por vizinhança gênica conservada (expansão fixa e dinâmica). Além de tipos diferentes de configurações de parâmetros para prever interações por perfil filogenético conservado. Optamos por explorar os métodos de vizinhança gênica conservada sem alterar as configurações escolhidas para o método de perfil filogenético conservado. O objetivo

foi padronizar as análises. A Tabela 10 mostra as configurações de parâmetros utilizadas na geração de redes de PPI com o nosso software.

Tabela 10 – Conjunto de parâmetros do GenPPI utilizados na geração de redes de interação para comparação com o STRING.

Id	Parâmetros
f1	genppi -expt fixed -w1 10 -cw1 3 -ppdiff tolerated 3 -ppiterlimit 1000000
f2	genppi -expt fixed -w1 10 -cw1 4 -trim 20000
f3	genppi -expt fixed -w1 10 -cw1 1 -ppiterlimit 500000
f4	genppi -expt fixed -w1 10 -cw1 1 -ppcomplete -aadiff limit 0 -aacheck limit 24
d1	genppi -expt dynamic -ws 10 -ppdiff tolerated 3 -ppiterlimit 1000000
d2	genppi -expt dynamic -ws 3 -trim 20000
d3	genppi -expt dynamic -ws 10 -ppiterlimit 500000
d4	genppi -expt dynamic -ws 10 -ppcomplete -aadiff limit 0 -aacheck limit 24
d5	genppi -expt dynamic -ws 3 -ppcomplete -ppdiff tolerated 1 -pphistofilter

Os prefixos f e d na coluna Id da Tabela 10, são referentes às análises com a expansão fixa e dinâmica, respectivamente. Uma explicação detalhada para os parâmetros utilizados, pode ser vista na Seção A.1 do Apêndice A (Guia do Usuário).

Os resultados das configurações de parâmetros utilizadas se resumem em redes de interação com diferentes números de nós (proteínas) e arestas (interações). Ao todo, foram geradas 9 redes de PPI para comparação com uma rede do STRING. Como dito na seção de métodos de avaliação, utilizamos duas abordagens de comparação: direta e indireta. Na sequência, são apresentadas essas duas abordagens.

4.2.4.1 Comparação Direta

Ao comparar diretamente redes de PPI, focamos em identificar interações de pares de proteínas em comum e específicas entre as redes comparadas. Para tanto, os conjuntos de pares de proteínas interagentes de cada rede, foram inseridos em tabelas distintas de um banco de dados. Depois, para identificar interações compartilhadas entre os conjuntos, verificou-se a intersecção par-a-par de tabelas por meio de uma consulta *sql*. Assim, obtivemos o percentual de interações em comum entre as redes comparadas.

Na Figura 38, consultamos quantos por cento das interações de uma rede listada numa coluna, encontram-se em redes referenciadas nas linhas. Como resultado, identificamos as interações em comum entre nossas redes e a RI do STRING. As redes criadas pelos Ids f1 e d1 são as mais numerosas. A razão é que os parâmetros que definimos nessas execuções do GenPPI, permitem explorar um número mais significativo de possibilidades. Justificamos tal conclusão porque todas as outras redes que geramos, e mesmo a RI do STRING, têm a maioria ou uma parte significativa de suas interações no conjunto das redes criadas pelos Ids f1 e d1 (células acinzentadas escuras e valores próximos a um). Também alcançamos a maior proximidade de nossas redes com a do STRING via Ids f1 e d1. O GENNPI previu quase a metade (46%) de todas as arestas de PPI preditas pelo STRING. Por

outro lado, o número mais significativo de nossas arestas previstas pelo STRING foi de 14%. Em relação aos resultados na Figura 38, o leitor deve notar que produzimos as redes do GenPPI usando apenas 50 genomas de *Corynebacterium*, enquanto o STRING usa um grupo muito maior, mais de cinco mil genomas, incluindo vários outros gêneros bacterianos. Considerando que usamos menos de 1% dos genomas hospedados pelo site STRING na geração de nossas redes de PPI, acreditamos que prever quase a metade das interações preditas pelo STRING, é um resultado bastante satisfatório. Também acreditamos que se o site STRING usar nosso conjunto de genomas, o mesmo poderia adquirir uma intersecção mais notável com nossas redes, mais significativa do que 14%.

N	Query column	902825	71492	278341	120647	914178	93373	295615	131356	280102	200088
Subject line	Test	f1	f2	f3	f4	d1	d2	d3	d4	d5	STRING
902825	f1	-	1.00	1.00	0.75	0.99	0.87	0.96	0.72	0.68	0.46
71492	f2	0.08	-	0.26	0.21	0.08	0.76	0.24	0.19	0.11	0.05
278341	f3	0.31	1.00	-	0.42	0.30	0.80	0.94	0.39	0.29	0.16
120647	f4	0.10	0.36	0.18	-	0.10	0.28	0.17	0.91	0.42	0.07
914178	d1	1.00	0.99	1.00	0.74	-	1.00	1.00	0.77	0.72	0.46
93373	d2	0.09	0.99	0.27	0.21	0.10	-	0.32	0.29	0.19	0.06
295615	d3	0.31	0.99	1.00	0.42	0.32	1.00	-	0.47	0.36	0.17
131356	d4	0.10	0.35	0.19	0.99	0.11	0.41	0.21	-	0.46	0.08
280102	d5	0.21	0.44	0.30	0.97	0.22	0.58	0.34	0.98	-	0.14
200088	STRING	0.10	0.14	0.12	0.11	0.10	0.14	0.12	0.11	0.10	-

Figura 38 – Gráfico de calor sobre o percentual de proteínas compartilhadas. Consultamos quantos por cento das interações de uma rede listada numa coluna, encontram-se em redes referenciadas nas linhas. O resultado compara as interações em comum entre diferentes redes do GenPPI e a rede do STRING. A intersecção mais significativa do GenPPI obteve 46% das interações do STRING (IDs f1 e d1). Por outro lado, o resultado mais significativo da rede do STRING em relação a nossas redes, atingiu 14%.

4.2.4.2 Comparação Indireta

Na comparação indireta o foco foi em características topológicas de redes complexas, que podem fornecer uma visão geral sobre a qualidade de uma rede de interação proteica,

para realização de análises biológicas. As características topológicas avaliadas são definidas a partir de métricas como: (i) número de nós, (ii) número de arestas, (iii) grau médio dos nós e (iv) densidade. Os valores apontados como satisfatórios para essas características, foram calibrados empiricamente. O significado das métricas i a iv no contexto de redes de interação entre proteínas, foi explicado na Subseção 4.1.4 referente ao método de comparação adotado. Os valores estatísticos dessas métricas referentes às redes de PPI comparadas, foram obtidos através do programa GEPHI (software para análise de topologia de grafos).

Utilizamos os critérios de i a iv como diretrizes para gerar redes com uma topologia adequada para análises biológicas. Assim, buscou-se obter redes com métricas mais aproximadas da rede do STRING que é nossa referência. Sobre a relevância biológica dos valores obtidos para as métricas utilizadas, tem-se que altos valores no número de nós simbolizam uma rede com muitas proteínas. Mais proteínas podem representar um organismo de uma melhor forma. Nesse sentido, são buscados valores aproximados ou até maiores em relação ao obtido pela RI do STRING. Tal busca inclui também a métrica de número de arestas (interações).

Outra questão biológica importante que se baseia nas características topológicas aqui estudadas, diz respeito às métricas de densidade e grau médio. No tocante a tais medidas, busca-se valores que não sejam maiores do que os encontrados na rede do STRING. Essa busca se baseia no fato de altas densidades representarem redes altamente conectadas, nas quais não se consegue inferir muitas informações. Baixos valores de grau médio simbolizam redes onde as interações entre proteínas se apresentam de uma forma mais distribuída proporcionando mais clareza para análises diversas.

Os resultados obtidos pelas execuções do programa, bem como os valores de métricas da rede gerada pelo STRING, encontram-se descritos na Tabela 11.

Tabela 11 – Valores de métricas da rede de interação do STRING em comparação com redes obtidas pelo GenPPI.

Id	Programa	Nº de nós	Nº de arestas	Grau médio	Densidade
0	STRING	2213	200088	180,83	0,082
f1	GenPPI	2149	902825	840,228	0,391
f2	GenPPI	2050	71492	69,748	0,034
f3	GenPPI	2057	278341	270,628	0,132
f4	GenPPI	1984	120647	121,620	0,061
d1	GenPPI	2141	914178	853,973	0,399
d2	GenPPI	2045	93373	91,318	0,045
d3	GenPPI	2045	295615	289,11	0,141
d4	GenPPI	1976	131356	132,947	0,067
d5	GenPPI	2058	280102	272,208	0,132

Os resultados das métricas da Tabela 11, referentes às redes de interação geradas por nossa ferramenta, foram obtidos através das configurações de parâmetros apresentadas na

Tabela 10. Em tais análises, pode-se perceber que as redes geradas pelo GenPPI possuem números de nós bem próximos da rede do STRING. Já no tocante ao número de arestas que representa a quantidade de interações preditas, observou-se valores tanto maiores (Ids f1, f3, d1, d3 e d5) quanto menores (Ids f2, f4, d2 e d4) do que os encontrados na rede de referência (Id 0).

O conjunto de parâmetros referente aos Ids f1 e d1 da Tabela 10, foram os responsáveis pelos maiores valores para quase todas as métricas. No entanto, como já foi explicado, há métricas que valores altos não significam boa topologia para análises numa rede de PPI, como o grau médio e densidade por exemplo. Para essas métricas obtivemos melhores resultados quando foram utilizados os conjuntos de parâmetros referentes aos Ids f2, f4, d2 e d4 da Tabela 10. Sobre a métrica de grau médio, as redes referentes a esses Ids na Tabela 11, apresentam valores iguais a 69,748, 121,62, 91,318 e 132,947. Estes representam, respectivamente, uma diferença de 61,42%, 32,74%, 49,50% e 26,47% a menos que o valor obtido pela rede do STRING. Valores menores para esse métrica podem significar redes onde as interações entre proteínas se apresentam de uma forma mais distribuída proporcionando mais clareza para análises diversas.

Em relação à métrica de densidade para a qual almeja-se valores menores, tendo em vista que altas densidades representam redes altamente conectadas nas quais não se consegue inferir muitas informações, apenas as configurações de parâmetros referentes aos Ids f2, f4, d2 e d4 da Tabela 10, é que possibilitaram um valor de densidade satisfatório. Ou seja, menor que 0,1, valor derivado empiricamente. Vale mencionar que a rede referente ao Id d4 do GenPPI, foi a que apresentou o valor mais próximo da rede do STRING para essa métrica, 0,067 versus 0,082, respectivamente.

4.2.4.3 Considerações Finais

Nas comparações realizadas, o nosso software identificou quase a metade (46%) de todas as arestas de PPI mapeadas pelo STRING. Por outro lado, o número mais significativo de arestas do STRING correspondentes nas nossas redes foi de 14%. Também constatamos que redes obtidas pelo GenPPI, podem apresentar valores de métricas aproximados dos da principal ferramenta do estado da arte. Assim, acredita-se que a solução proposta neste trabalho é capaz de gerar redes de interação de tão boas qualidades quanto as redes do STRING.

Nosso software transfere para o usuário, a decisão de quantos e quais genomas usar para embasar as predições. A chave para uma aplicação bem-sucedida do GenPPI, é a escolha de um conjunto adequado de genomas de referência. Isto é, um conjunto de genomas que não sejam evolutivamente muito distantes, e não muito próximos uns dos outros, de modo que apenas grupos de genes funcionalmente associados (proteínas interagentes) apresentem conservação gênica nos genomas de referência e sejam inferidas como PPIs. Por exemplo, se o objetivo do usuário é identificar as interações proteicas de uma espécie

específica, mas utiliza um conjunto de genomas do gênero como referência, então o programa não vai inferir a maior parte das interações dessa espécie, mas sim as do seu gênero bacteriano. Portanto, a principal limitação da nossa proposta é o desconhecimento, por parte do usuário, sobre as relações evolutivas dos genomas que se deseja analisar.

O maior desafio é encontrar o conjunto apropriado de espécies, a quantidade de genomas representantes de cada espécie e o conjunto adequado de parâmetros do programa, para gerar redes com características topológicas que permitam a investigação biológica que se deseja realizar. Por exemplo, gerar redes com baixos valores de densidade visando o estudo de medidas de centralidade com foco na identificação de alvos vacinais (HWANG; ZHANG; RAMANATHAN, 2008). Não há fórmula fechada para nenhuma dessas três variáveis. O usuário precisa explorar conjuntos de genomas de referência, conjuntos de parâmetros, e precisa executar o programa N vezes até concluir que obteve uma rede de interação adequada para as análises biológicas que se deseja realizar. Não há critérios cientificamente estabelecidos para julgar se uma rede é adequada ou não. As únicas características que se consegue encontrar para definir essa adequação, são: a quantidade de proteínas (número de nós), a quantidade total de interações entre essas proteínas (número de arestas), o grau médio e a densidade.

4.2.5 Exemplo de uma Rede de Interação Proteína-Proteína Gerada Pelo GenPPI

A título de demonstração, a Figura 39 mostra a montagem visual feita através do GEPHI para uma rede de PPI de *Corynebacterium pseudotuberculosis*. Essa RI se refere a uma das 50 espécies de *Corynebacterium* computadas neste trabalho. Na Figura 39, as cores de vértices (proteínas), representam os locais subcelulares dessas proteínas, previstos com o software SurfG Plus (BARINOV et al., 2009). Os vértices de maior tamanho são nós destacados com a métrica Bridging Centrality capaz de identificar proteínas que servem como ponte entre aglomerados de proteínas interagentes em uma rede de interação. Em teoria, se desligarmos os genes que são pontes entre processos bioquímicos (aglomerados de proteínas interagentes) responsáveis pela infecção no hospedeiro, então poderíamos neutralizar um processo infeccioso. Portanto, acredita-se que tais proteínas destacadas em uma RI, através da métrica Bridging Centrality, podem ser úteis na confecção de drogas e vacinas contra as doenças causadas por organismos bacterianos (HWANG; ZHANG; RAMANATHAN, 2008). Na rede de interação da *Corynebacterium pseudotuberculosis strain:Ft_2193/67* ilustrada pela Figura 39, essas proteínas foram destacadas por meio de um plugin desenvolvido para o software GEPHI pela equipe do professor que orientou o desenvolvimento deste trabalho. Esse plugin aplica a métrica bridging centrality que permite identificar nós que interligam componentes densamente conectados de uma rede complexa.

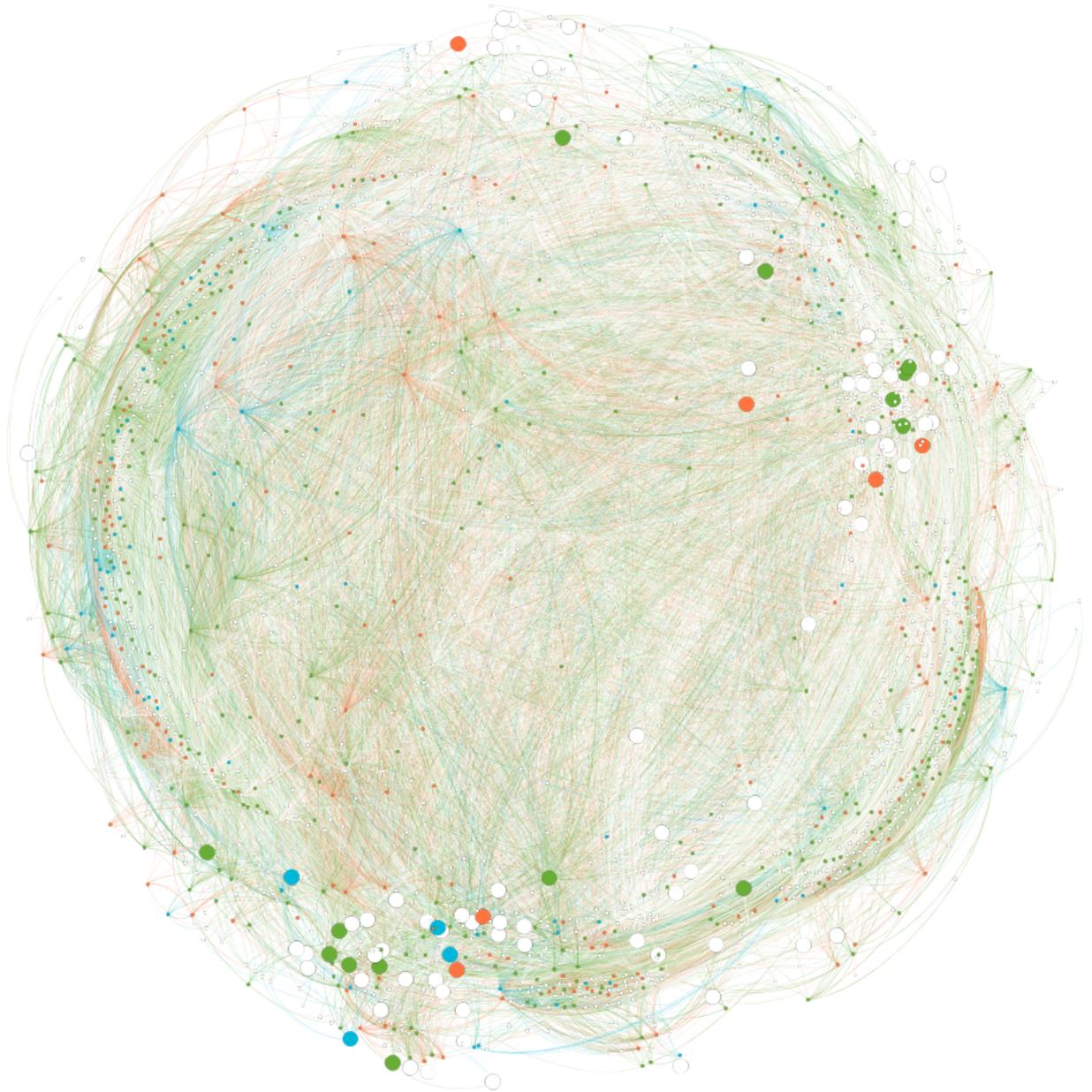


Figura 39 – Uma rede de interações proteicas gerada pelo GenPPI para o genoma *Corynebacterium pseudotuberculosis strain:Ft_2193/67*. Os locais subcelulares das proteínas (vértices), representados pelo esquema de cores adotado, são o citoplasma (branco), a membrana (verde), potencialmente expostas na superfície (laranja) e secretadas (azul). Vértices maiores se destacam de acordo com a métrica Bridging Centrality. Esta rede foi desenhada pelo software GEPHI executando os algoritmos de distribuição de dados, nesta ordem: Yifan Hu Multilevel, Fruchterman-Reingold e layouts Force Atlas.

Conclusão

Este trabalho teve como objetivo geral contribuir no campo da biologia computacional fornecendo um novo software autônomo de bioinformática para predição *ab initio* de redes de interação entre proteínas bacterianas. A partir de dados gerados por essa ferramenta denominada GenPPI, foi possível representar com alto grau de confiança, as relações evolutivas de espécies e subespécies de gêneros bacterianos selecionados para estudos de caso. Sustentamos a hipótese de correção e confiabilidade biológica para nossas predições computacionais, com base na qualidade de dados gerados pelo programa, para inferir interações de proteínas. Visto que gráficos elaborados a partir desses dados, refletiram com exatidão, o conhecimento literário sobre espécies e subespécies de gêneros bacterianos analisados, concluí-se que redes de PPI preditas pelo software proposto, são biologicamente corretas e confiáveis.

O método implementado faz predições de PPI a partir de evidências biológicas de interação proteica, encontradas nos próprios genomas de interesse do usuário. À vista disso, acredita-se que o GenPPI supriu uma deficiência identificada no estado da arte, a indisponibilidade de soluções computacionais para prever interações sem negligenciar as proteínas inéditas de novos genomas elucidados. Proteínas essas que representam pelo menos 10% dos genes desses genomas (LAPIERRE; GOGARTEN, 2009).

O objetivo geral deste trabalho foi alcançado progressivamente, na medida que atingiu-se cada objetivo específico listado na Subseção 1.2.3.

O primeiro objetivo específico é referente ao desenvolvimento do software proposto. Essa tarefa foi apresentada com detalhes no Capítulo 3 discorrendo inicialmente sobre as tecnologias utilizadas no desenvolvimento da proposta (Seção 3.2) e, posteriormente, sobre as etapas de implementação da mesma.

O segundo objetivo específico listado na Subseção 1.2.3, diz respeito às análises computacionais realizadas com diversos organismos de importância relevante para a medicina e economia. Esse objetivo foi atingido por meio de 3 estudos de caso: (i) 28 genomas do gênero *Dietzia* e 45 de *Rhodococcus*, (ii) 50 de *Corynebacterium* e (iii) 81 de *Aeromonas*, foram analisados pelo software desenvolvido. Como fruto de tais análises, destaca-se

uma contribuição feita ao grupo RECOM que utilizou dados gerados pelo GenPPI para classificar novas espécies de bactérias do gênero *Aeromonas*.

O terceiro objetivo específico listado na Subseção 1.2.3, é um objetivo de importância crucial se referindo à validação biológica das predições computacionais realizadas. Tal validação foi feita durante os três estudos de casos realizados, utilizando dados gerados pelo programa para comparar estatisticamente espécies evolutivamente correlacionadas. A expectativa dessa comparação foi que genomas de espécies conhecidas na literatura como evolutivamente próximas, apresentassem medidas estatísticas muito similares e vice-versa. Através dos gráficos e avaliação de resultados da Subseção 4.2.3, foram apresentadas várias evidências nesse sentido para genomas dos três gêneros bacterianos selecionados como estudos de caso. Portanto, a expectativa acima mencionada foi atingida validando assim, a hipótese de que a solução computacional proposta seria capaz de prever redes de PPI biologicamente corretas.

Por final, para atingir o quarto e último objetivo específico listado na Subseção 1.2.3, foram utilizadas várias configurações de parâmetros do programa, para gerar de redes de interação do genoma de *Corynebacterium pseudotuberculosis strain:Ft_2193/67*. O objetivo foi estimar a qualidade das redes geradas pelo software desenvolvido, frente a uma rede desse mesmo organismo gerada pela principal ferramenta do estado da arte (STRING). Mostramos que a solução proposta neste trabalho é capaz de gerar redes de interação de tão boas qualidades quanto as redes do STRING. A principal evidência é que o nosso software previu quase a metade (46%) de todas as arestas de PPI preditas pelo STRING, enquanto o número mais significativo de nossas arestas previstas pelo STRING foi de 14%. Também constatamos que redes obtidas pelo GenPPI, podem apresentar características topológicas com valores aproximados dos da principal ferramenta do estado da arte.

Posto que todos os objetivos específicos foram atingidos, entende-se como cumprido o objetivo geral deste trabalho. Isto é, contribuir no campo da biologia computacional fornecendo um novo software autônomo de bioinformática para predição *ab initio* de redes de interação entre proteínas bacterianas.

As hipóteses levantadas foram: possibilidade de fazer predições de redes de PPI biologicamente corretas e confiáveis com a ferramenta proposta (validada na Subseção 4.2.3); realizar predições *ab initio* em um tempo aceitável (validada na Subseção 4.2.1) e a possibilidade de gerar redes de interações proteicas com níveis de qualidade e confiabilidade próximos do que oferece a principal ferramenta do estado da arte (STRING) (validada na Subseção 4.2.4).

Depois de concluir que os objetivos específicos e geral foram alcançados, as próximas seções tratam de destacar as principais contribuições e apresentar os trabalhos futuros para melhoria da proposta atual.

5.1 Principais Contribuições

Este trabalho de pesquisa contribui sob diversos aspectos que englobam desde a disponibilização de uma nova ferramenta de bioinformática, até a geração de dados e informações inéditas sobre organismos bacterianos. Na sequência estão destacadas as principais contribuições deste trabalho.

- Disponibilização de um novo software autônomo de bioinformática para predição *ab initio* de redes de interação entre proteínas bacterianas, que está disponível para download no site: <<https://genppi.facom.ufu.br/>> ou no repositório: <<https://github.com/santosardr/genppi>>. Nesse repositório também contém um guia do usuário para instruí-lo sobre a utilização do programa. A hipótese de que nossas predições computacionais seriam biologicamente corretas e confiáveis, foi validada com base na construção de mapas de filogenia finamente detalhados para os genomas estudados. Tal validação foi feita ao longo da Subseção 4.2.3 onde mostramos que através de dados gerados pelo software proposto, pode-se distinguir, por exemplo, entre biovars (subespécies) da espécie *Corynebacterium pseudotuberculosis* (BERNARDES et al., 2020). Além da separação ideal entre organismos procariotos de outros gêneros bacterianos. Considerando a qualidade de separação das espécies estudadas, demonstrada em gráficos feitos a partir dos dados que sustentam as predições do GenPPI, reivindica-se boa qualidade e confiabilidade para redes de PPI geradas pelo nosso software, além de menos recursos computacionais necessários para tal tarefa. Por exemplo, conforme apresentado na Subseção 4.2.1, foi necessário menos de uma hora de processamento para computar 50 genomas de *Corynebacterium pseudotuberculosis* contendo em média 2.200 proteínas cada. Isso valida outra hipótese levantada inicialmente: a possibilidade de realizar predições *ab initio* de interações proteicas em um tempo aceitável. Vale mencionar que foi criada uma página na internet (<<https://genppi.facom.ufu.br/>>) para que o usuário possa fazer o upload de relatórios gerados automaticamente em uma execução do programa (A.2), e criar gráficos como os que foram apresentados na Subseção 4.2.3. Tais gráficos são úteis para fazer análises de separação de espécies.

- O nosso software inspeciona genomas representados com arquivos multi-fasta de proteínas, procurando por eventos de vizinhança gênica conservada, perfil filogenético e fusão gênica para inferir interações entre proteínas. Permite transferir para o usuário final, a decisão de quantos e quais genomas de referência usar para embasar as predições e construir redes de interação de proteínas. Isso permite ao usuário explorar diversos conjuntos de genomas de referência visando produzir redes de interação mais adequadas para uma demanda de pesquisa específica. Destaca-se essa flexibilidade como sendo um diferencial do GenPPI em relação às alternativas do

estado da arte. Além disso, com o software proposto, é suprida uma deficiência identificada no estado da arte, a indisponibilidade de alternativas computacionais para prever PPIs sem negligenciar proteínas inéditas.

- No contexto de estudos de filogenia, contribuimos com a rede de pesquisa RECOM, na distinção de espécies inéditas do gênero bacteriano *Aeromonas*, recentemente sequenciadas e montadas por esse grupo. Alguns resultados e conclusões das análises computacionais realizadas para essas espécies (Subseção 4.2.3.3), foram incluídos em um artigo que, até o mês de Dezembro de 2020, ainda estava sendo produzido por essa rede de pesquisa. É importante salientar que as nossas análises computacionais foram decisivas nas conclusões desse artigo sob redação. O motivo é que os métodos bioquímicos comumente utilizados para diferenciar bactérias, não possuem exatidão suficiente para divisar entre espécies e subespécies do gênero bacteriano *Aeromonas*, o que foi possível através do GenPPI.
- Este trabalho também contribui introduzindo uma nova heurística para comparação par-a-par de sequências de aminoácidos de proteínas. Tal heurística é uma alternativa viável frente a utilização de um algoritmo exato ou até mesmo das melhores heurísticas atuais para comparação par-a-par de sequências de proteínas. A eficácia da heurística proposta foi demonstrada pela análise de resultados dos experimentos descritos na Subseção 4.2.2. Comparamos a nossa heurística com o principal algoritmo heurístico do estado da arte, o BLASTp (ALTSCHUL et al., 1990). A exatidão dos dois algoritmos heurísticos foi estimada verificando qual se aproxima mais do algoritmo exato de alinhamento de sequências biológicas Needleman-Wunsch (NEEDLEMAN; WUNSCH, 1970). Verificou-se que a heurística proposta superou o BLASTp em termos de tempo de processamento (principalmente) e em exatidão na comparação par-a-par de proteínas.

5.2 Trabalhos Futuros

O GenPPI faz suas predições de pares de proteínas interagentes, a partir de dados genômicos baseados nos eventos evolutivos das espécies. Eventos como vizinhança gênica conservada, fusão gênica e perfil filogenético conservado, são características comumente utilizadas na inferência de PPIs a partir de dados genômicos. Além desses três métodos, existem outros que também podem ser utilizados para prever interações entre proteínas bacterianas. Nesse sentido, como trabalhos futuros para melhoria da proposta atual, destaca-se a exploração de outros métodos de inferência de PPIs a partir de genomas, como os apresentados em (VALENCIA; PAZOS, 2002), por exemplo.

Diante dessa possibilidade, acredita-se ser possível o enriquecimento de nossas análises computacionais e a melhoria da proposta atual.

Como trabalhos futuros destaca-se também o desenvolvimento de outro software autônomo baseado na heurística do GenPPI. O objetivo desse novo software (Histo-Fasta) seria fazer comparação par-a-par de sequências de proteínas e também geração rápida de pan-genomas. A heurística proposta neste trabalho apresentou resultados melhores em relação à principal alternativa de comparação de sequências proteicas do estado da arte (BLASTp). Essa melhoria é refletida significativamente pelo tempo de processamento e marginalmente pela exatidão na comparação de sequências proteicas. Portanto, acredita-se que essa proposta de trabalho futuro, é uma alternativa viável frente as opções atuais para comparação par-a-par de proteínas, ou frente as opções de geração rápida de pan-genomas, como o software Roary por exemplo (PAGE et al., 2015).

5.3 Contribuições em Produção Bibliográfica

Submissão como autor principal do artigo "*GenPPI: standalone software for creating protein interaction networks from genomes*" no periódico internacional **BMC Bioinformatics** (Qualis/CAPES A1), em maio de 2021. Esse artigo conta com resultados apresentados nas Subseções 4.2.2.1, 4.2.3.2 e 4.2.4.

Referências

- ALTSCHUL, S. F. et al. Basic local alignment search tool. **Journal of molecular biology**, Elsevier, v. 215, n. 3, p. 403–410, 1990. Disponível em: <[https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)>.
- AMBROGELLY, A.; PALIOURA, S.; SÖLL, D. Natural expansion of the genetic code. **Nature chemical biology**, Nature Publishing Group, v. 3, n. 1, p. 29–35, 2007. Disponível em: <<https://www.nature.com/articles/nchembio847>>.
- ANANTHASUBRAMANIAN, S. et al. Mycobacterium tuberculosis and clostridium difficile interactomes: demonstration of rapid development of computational system for bacterial interactome prediction. **Microbial informatics and experimentation**, Springer, v. 2, n. 1, p. 4, 2012. Disponível em: <<https://doi.org/10.1186/2042-5783-2-4>>.
- ANCESTRYDNA. **The DNA Code and Codons**. 2020. Disponível em: <<https://www.ancestry.com/lp/dna-sequencing/dna-code-codons>>.
- ARKIN, M. R.; WELLS, J. A. Small-molecule inhibitors of protein–protein interactions: progressing towards the dream. **Nature reviews Drug discovery**, Nature Publishing Group, v. 3, n. 4, p. 301–317, 2004. Disponível em: <<https://doi.org/10.1038/nrd1343>>.
- BADELL, E. et al. *Corynebacterium rouxii* sp. nov., a novel member of the diphtheriae species complex. **Research in Microbiology**, Elsevier, 2020. Disponível em: <<https://doi.org/10.1016/j.resmic.2020.02.003>>.
- BADER, G. D.; HOGUE, C. W. An automated method for finding molecular complexes in large protein interaction networks. **BMC bioinformatics**, Springer, v. 4, n. 1, p. 2, 2003. Disponível em: <<https://doi.org/10.1186/1471-2105-4-2>>.
- BARINOV, A. et al. Prediction of surface exposed proteins in streptococcus pyogenes, with a potential application to other gram-positive bacteria. **Proteomics**, Wiley Online Library, v. 9, n. 1, p. 61–73, 2009. Disponível em: <<https://doi.org/10.1002/pmic.200800195>>.
- BERNARDES, J. S. et al. A comparative pan-genomic analysis of 53 c. pseudotuberculosis strains based on functional domains. **Journal of Biomolecular Structure and Dynamics**, Taylor & Francis, p. 1–13, 2020. Disponível em: <<https://doi.org/10.1080/07391102.2020.1805017>>.

- BONETTI, D. R. F. **Algoritmos de estimação de distribuição para predição ab initio de estruturas de proteínas**. Tese (Doutorado) — Universidade de São Paulo, 2012. Disponível em: <<https://www.teses.usp.br/teses/disponiveis/55/55134/tde-03082015-193613/en.php>>.
- BRAUN, P. Interactome mapping for analysis of complex phenotypes: insights from benchmarking binary interaction assays. **Proteomics**, Wiley Online Library, v. 12, n. 10, p. 1499–1518, 2012. Disponível em: <<https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/pmic.201100598>>.
- BROWNE, F. et al. From experimental approaches to computational techniques: a review on the prediction of protein-protein interactions. **Advances in Artificial Intelligence**, Hindawi Limited, v. 2010, 2010. Disponível em: <<https://doi:10.1155/2010/924529>>.
- BUSCH, A. et al. Genome sequence of a pathogenic corynebacterium ulcerans strain isolated from a wild boar with necrotizing lymphadenitis. **BMC research notes**, BioMed Central, v. 12, n. 1, p. 1–3, 2019. Disponível em: <<https://doi.org/10.1186/s13104-019-4704-3>>.
- CHANDRASHEKAR, D. S. et al. Ualcan: a portal for facilitating tumor subgroup gene expression and survival analyses. **Neoplasia**, Elsevier, v. 19, n. 8, p. 649–658, 2017. Disponível em: <<https://doi.org/10.1016/j.neo.2017.05.002>>.
- CHAREST, A. et al. Fusion of fig to the receptor tyrosine kinase ros in a glioblastoma with an interstitial del (6)(q21q21). **Genes, Chromosomes and Cancer**, Wiley Online Library, v. 37, n. 1, p. 58–71, 2003. Disponível em: <<https://onlinelibrary.wiley.com/doi/abs/10.1002/gcc.10207>>.
- CHEN, G.; WANG, X.; LI, X. **Fundamentals of complex networks: models, structures and dynamics**. John Wiley & Sons, 2014. Disponível em: <<https://doi.org/10.1002/9781118718124>>.
- CHEN, J. Y.; PANDEY, R.; NGUYEN, T. M. Happi-2: a comprehensive and high-quality map of human annotated and predicted protein interactions. **BMC genomics**, Springer, v. 18, n. 1, p. 182, 2017. Disponível em: <<https://doi.org/10.1186/s12864-017-3512-1>>.
- CHISNALL, D. **Cocoa Programming Developer's Handbook**. [S.l.]: Addison-Wesley Professional, 2009.
- CLAVERYS, J.-P. et al. Construction and evaluation of new drug-resistance cassettes for gene disruption mutagenesis in streptococcus pneumoniae, using an ami test platform. **Gene**, Elsevier, v. 164, n. 1, p. 123–128, 1995. Disponível em: <[https://doi.org/10.1016/0378-1119\(95\)00485-O](https://doi.org/10.1016/0378-1119(95)00485-O)>.
- CLOUGH, E.; BARRETT, T. The gene expression omnibus database. In: **Statistical genomics**. Springer, 2016. p. 93–110. Disponível em: <https://doi.org/10.1007/978-1-4939-3578-9_5>.
- CONSORTIUM, G. O. The gene ontology resource: 20 years and still going strong. **Nucleic acids research**, Oxford University Press, v. 47, n. D1, p. D330–D338, 2019. Disponível em: <<https://doi.org/10.1093/nar/gky1055>>.

CRICK, F. Central dogma of molecular biology. **Nature**, Nature Publishing Group, v. 227, n. 5258, p. 561–563, 1970. Disponível em: <<https://doi.org/10.1038/227561a0>>.

DEY, L.; MUKHOPADHYAY, A. A classification-based approach to prediction of dengue virus and human protein-protein interactions using amino acid composition and conjoint triad features. In: IEEE. **2019 IEEE Region 10 Symposium (TENSYP)**. 2019. p. 373–378. Disponível em: <<https://ieeexplore.ieee.org/abstract/document/8971382>>.

DURBIN, R. et al. **Biological sequence analysis: probabilistic models of proteins and nucleic acids**. [S.l.]: Cambridge university press, 1998.

DYER, M. D.; MURALI, T.; SOBRAL, B. W. The landscape of human proteins interacting with viruses and other pathogens. **PLoS Pathog**, Public Library of Science, v. 4, n. 2, p. e32, 2008. Disponível em: <<https://doi.org/10.1371/journal.ppat.0040032>>.

ENRIGHT, A. J. et al. Protein interaction maps for complete genomes based on gene fusion events. **Nature**, Nature Publishing Group, v. 402, n. 6757, p. 86–90, 1999. Disponível em: <<https://doi.org/10.1038/47056>>.

ESCH, R.; MERKL, R. Conserved genomic neighborhood is a strong but no perfect indicator for a direct interaction of microbial gene products. **BMC bioinformatics**, Springer, v. 21, n. 1, p. 1–8, 2020. Disponível em: <<https://doi.org/10.1186/s12859-019-3200-z>>.

GUARALDI, A. L. de M.; HIRATA, R.; AZEVEDO, V. A. de C. Corynebacterium diphtheriae, corynebacterium ulcerans and corynebacterium pseudotuberculosis—general aspects. In: **Corynebacterium diphtheriae and Related Toxigenic Species**. Springer, 2014. p. 15–37. Disponível em: <https://link.springer.com/chapter/10.1007/978-94-007-7624-1_2>.

HE, F. et al. The prediction of protein-protein interaction networks in rice blast fungus. **BMC genomics**, Springer, v. 9, n. 1, p. 519, 2008. Disponível em: <<https://doi.org/10.1186/1471-2164-9-519>>.

HIROSE, S. Inferring protein-protein interactions (ppis) based on computational methods. **Protein-Protein Interactions: Computational and Experimental Tools**, BoD—Books on Demand, p. 147, 2012. Disponível em: <<https://www.intechopen.com/books/protein-protein-interactions-computational-and-experimental-tools/inferring-protein-protein-interactions-ppis-based-on-computational-methods>>.

HOU, J. **New approaches of protein function prediction from protein interaction networks**. Academic Press, 2017. Disponível em: <<https://www.sciencedirect.com/book/9780128098141/new-approaches-of-protein-function-prediction-from-protein-interaction-networks>>.

HOYTE, D. **Let over lambda**. Lulu. com, 2008. Disponível em: <<https://dl.acm.org/doi/book/10.5555/1816935>>.

HWANG, W.-C.; ZHANG, A.; RAMANATHAN, M. Identification of information flow-modulating drug targets: a novel bridging paradigm for drug discovery. **Clinical Pharmacology & Therapeutics**, Wiley Online Library, v. 84, n. 5, p. 563–572, 2008. Disponível em: <<https://doi.org/10.1038/clpt.2008.129>>.

- JEONG, H. et al. Lethality and centrality in protein networks. **Nature**, Nature Publishing Group, v. 411, n. 6833, p. 41–42, 2001. Disponível em: <<https://doi.org/10.1038/35075138>>.
- JIAO, X. et al. David-ws: a stateful web service to facilitate gene/protein list analysis. **Bioinformatics**, Oxford University Press, v. 28, n. 13, p. 1805–1806, 2012. Disponível em: <<https://doi.org/10.1093/bioinformatics/bts251>>.
- KANEHISA, M. et al. New approach for understanding genome variations in kegg. **Nucleic acids research**, Oxford University Press, v. 47, n. D1, p. D590–D595, 2019. Disponível em: <<https://doi.org/10.1093/nar/gky962>>.
- KESKIN, O. et al. Principles of protein- protein interactions: What are the preferred ways for proteins to interact? **Chemical reviews**, ACS Publications, v. 108, n. 4, p. 1225–1244, 2008. Disponível em: <<https://doi.org/10.1021/cr040409x>>.
- KOUTROULI, M. et al. A guide to conquer the biological network era using graph theory. **Frontiers in Bioengineering and Biotechnology**, Frontiers, v. 8, p. 34, 2020. Disponível em: <<https://doi.org/10.3389/fbioe.2020.00034>>.
- LAPIERRE, P.; GOGARTEN, J. P. Estimating the size of the bacterial pan-genome. **Trends in genetics**, Elsevier, v. 25, n. 3, p. 107–110, 2009. Disponível em: <<https://doi.org/10.1016/j.tig.2008.12.004>>.
- LEONARD, M.; GRAHAM, S.; BONACUM, D. The human factor: the critical importance of effective teamwork and communication in providing safe care. **BMJ Quality & Safety**, BMJ Publishing Group Ltd, v. 13, n. suppl 1, p. i85–i90, 2004. Disponível em: <<http://dx.doi.org/10.1136/qshc.2004.010033>>.
- LISPCOOKBOOK. **The Common Lisp Cookbook – Performance Tuning and Tips**. 2021. Disponível em: <<https://lispcookbook.github.io/cl-cookbook/performance.html>>.
- LIU, Z.-P. et al. Inferring a protein interaction map of mycobacterium tuberculosis based on sequences and interologs. In: SPRINGER. **BMC bioinformatics**. 2012. v. 13, n. S7, p. S6. Disponível em: <<https://doi.org/10.1186/1471-2105-13-S7-S6>>.
- LOPES, C. T. et al. Cytoscape web: an interactive web-based network browser. **Bioinformatics**, Oxford University Press, v. 26, n. 18, p. 2347–2348, 2010. Disponível em: <<https://doi.org/10.1093/bioinformatics/btq430>>.
- MERING, C. v. et al. String: a database of predicted functional associations between proteins. **Nucleic acids research**, Oxford University Press, v. 31, n. 1, p. 258–261, 2003. Disponível em: <<https://doi.org/10.1093/nar/gkg034>>.
- MOSTAFAVI, S.; MORRIS, Q. Combining many interaction networks to predict gene function and analyze gene lists. **Proteomics**, Wiley Online Library, v. 12, n. 10, p. 1687–1696, 2012. Disponível em: <<https://doi.org/10.1002/pmic.201100607>>.
- NEEDLEMAN, S. B.; WUNSCH, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. **Journal of molecular biology**, Elsevier, v. 48, n. 3, p. 443–453, 1970. Disponível em: <[https://doi.org/10.1016/0022-2836\(70\)90057-4](https://doi.org/10.1016/0022-2836(70)90057-4)>.

- NEWMAN, M. E. Estimating network structure from unreliable measurements. **Physical Review E**, APS, v. 98, n. 6, p. 062321, 2018. Disponível em: <<https://doi.org/10.1103/PhysRevE.98.062321>>.
- NG, K. L.; KHOR, S. M. Graphite-based nanocomposite electrochemical sensor for multiplex detection of adenine, guanine, thymine, and cytosine: A biomedical prospect for studying dna damage. **Analytical chemistry**, ACS Publications, v. 89, n. 18, p. 10004–10012, 2017. Disponível em: <<https://doi.org/10.1021/acs.analchem.7b02432>>.
- NIWA, H. et al. Characterization of human clinical isolates of dietzia species previously misidentified as rhodococcus equi. **European journal of clinical microbiology & infectious diseases**, Springer, v. 31, n. 5, p. 811–820, 2012. Disponível em: <<https://doi.org/10.1007/s10096-011-1379-7>>.
- NORVIG, P. **Paradigms of artificial intelligence programming: case studies in Common LISP**. Morgan Kaufmann, 1992. Disponível em: <<https://www.elsevier.com/books/paradigms-of-artificial-intelligence-programming/norvig/978-0-08-057115-7>>.
- PAGE, A. J. et al. Roary: rapid large-scale prokaryote pan genome analysis. **Bioinformatics**, Oxford University Press, v. 31, n. 22, p. 3691–3693, 2015. Disponível em: <<https://doi.org/10.1093/bioinformatics/btv421>>.
- PANCHENKO, A.; PRZYTYCKA, T. M. **Protein-protein interactions and networks: identification, computer analysis, and prediction**. Springer, 2010. v. 9. Disponível em: <<https://link.springer.com/book/10.1007%2F978-1-84800-125-1>>.
- PEARL, J. Intelligent search strategies for computer problem solving. **Ad-dision Wesley**, 1984. Disponível em: <<https://www.semanticscholar.org/paper/Intelligent-Search-Strategies-for-Computer-Problem-Pearl/2a2d5c2532cd2d1cf2b612a50baa2bb2fe1b5735?p2df>>.
- PELLEGRINI, M. et al. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 96, n. 8, p. 4285–4288, 1999. Disponível em: <<https://doi.org/10.1073/pnas.96.8.4285>>.
- PENG, X. et al. Protein–protein interactions: detection, reliability assessment and applications. **Briefings in bioinformatics**, Oxford University Press, v. 18, n. 5, p. 798–819, 2017. Disponível em: <<https://academic.oup.com/bib/article/18/5/798/2562794>>.
- PILARES, L. et al. Identification of atypical rhodococcus-like clinical isolates as dietzia spp. by 16s rrna gene sequencing. **Journal of clinical microbiology**, Am Soc Microbiol, v. 48, n. 5, p. 1904–1907, 2010. Disponível em: <<https://jcm.asm.org/content/48/5/1904.short>>.
- PRASAD, T. K. et al. Human protein reference database—2009 update. **Nucleic acids research**, Oxford University Press, v. 37, n. suppl_1, p. D767–D772, 2009. Disponível em: <<https://doi.org/10.1093/nar/gkn892>>.
- PUJANA, M. A. et al. Network modeling links breast cancer susceptibility and centrosome dysfunction. **Nature genetics**, Nature Publishing Group, v. 39, n. 11, p. 1338–1349, 2007. Disponível em: <<https://doi.org/10.1038/ng.2007.2>>.

- QI, Y. et al. Semi-supervised multi-task learning for predicting interactions between hiv-1 and human proteins. **Bioinformatics**, Oxford University Press, v. 26, n. 18, p. i645–i652, 2010. Disponível em: <<https://doi.org/10.1093/bioinformatics/btq394>>.
- RAMAN, K. Construction and analysis of protein–protein interaction networks. **Automated experimentation**, Springer, v. 2, n. 1, p. 2, 2010. Disponível em: <<https://doi.org/10.1186/1759-4499-2-2>>.
- RAMANATHAN, M.; PORTER, D. F.; KHAVARI, P. A. Methods to study rna–protein interactions. **Nature methods**, Nature Publishing Group, v. 16, n. 3, p. 225–234, 2019. Disponível em: <<https://doi.org/10.1038/s41592-019-0330-1>>.
- REZENDE, A. M. Predição computacional de interações de proteína-proteína em proteomas preditos de leishmania. Universidade Federal de Minas Gerais, 2012. Disponível em: <<https://repositorio.ufmg.br/handle/1843/BUOS-998GUB>>.
- ROMERO-MOLINA, S. et al. Ppi-detect: A support vector machine model for sequence-based prediction of protein–protein interactions. **Journal of computational chemistry**, Wiley Online Library, v. 40, n. 11, p. 1233–1242, 2019. Disponível em: <<https://doi.org/10.1002/jcc.25780>>.
- ROSSUM, G. V.; DRAKE, F. L. Python 2.6 reference manual. CreateSpace, 2009. Disponível em: <<https://dl.acm.org/doi/book/10.5555/1610526>>.
- SANCHEZ, C. et al. Grasping at molecular interactions and genetic networks in drosophila melanogaster using flynets, an internet database. **Nucleic acids research**, Oxford University Press, v. 27, n. 1, p. 89–94, 1999. Disponível em: <<https://doi.org/10.1093/nar/27.1.89>>.
- SEIBEL, P. **Practical common lisp**. Apress, 2006. Disponível em: <<http://www.gigamonkeys.com/book/>>.
- SHOEMAKER, B. A.; PANCHENKO, A. R. Deciphering protein–protein interactions. part i. experimental techniques and databases. **PLoS Comput Biol**, Public Library of Science, v. 3, n. 3, p. e42, 2007. Disponível em: <<https://doi.org/10.1371/journal.pcbi.0030042>>.
- SINGH, G. B. **Fundamental of bioinformatics and computational biology**. Springer, 2015. Disponível em: <<https://link.springer.com/book/10.1007%2F978-3-319-11403-3>>.
- SNEL, B.; BORK, P.; HUYNEN, M. Genome evolution: gene fusion versus gene fission. **Trends in genetics**, Elsevier, v. 16, n. 1, p. 9–11, 2000. Disponível em: <[https://doi.org/10.1016/S0168-9525\(99\)01924-1](https://doi.org/10.1016/S0168-9525(99)01924-1)>.
- SNEL, B. et al. String: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. **Nucleic acids research**, Oxford University Press, v. 28, n. 18, p. 3442–3444, 2000. Disponível em: <<https://doi.org/10.1093/nar/28.18.3442>>.
- STARK, C. et al. Biogrid: a general repository for interaction datasets. **Nucleic acids research**, Oxford University Press, v. 34, n. suppl_1, p. D535–D539, 2006. Disponível em: <<https://doi.org/10.1093/nar/gkj109>>.

- SUN, J.; LI, Y.; ZHAO, Z. Phylogenetic profiles for the prediction of protein–protein interactions: how to select reference organisms? **Biochemical and Biophysical Research Communications**, Elsevier, v. 353, n. 4, p. 985–991, 2007. Disponível em: <<https://doi.org/10.1016/j.bbrc.2006.12.146>>.
- SUN, Y.; ZHANG, Z. In silico identification of crucial genes and specific pathways in hepatocellular cancer. **Genetic Testing and Molecular Biomarkers**, Mary Ann Liebert, Inc., publishers 140 Huguenot Street, 3rd Floor New . . . , v. 24, n. 5, p. 296–308, 2020. Disponível em: <<https://doi.org/10.1089/gtmb.2019.0242>>.
- SZKLARCZYK, D. et al. String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. **Nucleic acids research**, Oxford University Press, v. 47, n. D1, p. D607–D613, 2019. Disponível em: <<https://doi.org/10.1093/nar/gky1131>>.
- TANG, Z. et al. Gepia: a web server for cancer and normal gene expression profiling and interactive analyses. **Nucleic acids research**, Oxford University Press, v. 45, n. W1, p. W98–W102, 2017. Disponível em: <<https://doi.org/10.1093/nar/gkx247>>.
- TEAM, R. C. et al. **R: A language and environment for statistical computing**. Vienna, Austria, 2013. Disponível em: <<http://cran.univ-paris1.fr/web/packages/dplR/vignettes/intro-dplR.pdf>>.
- VALENCIA, A.; PAZOS, F. Computational methods for the prediction of protein interactions. **Current opinion in structural biology**, Elsevier, v. 12, n. 3, p. 368–373, 2002. Disponível em: <[https://doi.org/10.1016/S0959-440X\(02\)00333-0](https://doi.org/10.1016/S0959-440X(02)00333-0)>.
- VERNIKOS, G. et al. Ten years of pan-genome analyses. **Current opinion in microbiology**, Elsevier, v. 23, p. 148–154, 2015. Disponível em: <<https://doi.org/10.1016/j.mib.2014.11.016>>.
- WANG, Y. et al. A survey of current trends in computational predictions of protein–protein interactions. **Frontiers of Computer Science**, Springer, v. 14, n. 4, p. 144901, 2020. Disponível em: <<https://doi.org/10.1007/s11704-019-8232-z>>.
- WHEATGENOMES. **Pan Genome**. 2020. Disponível em: <<http://www.10wheatgenomes.com/what-is-a-pan-genome>>.
- YANG, S. et al. Understanding human-virus protein-protein interactions using a human protein complex-based analysis framework. **MSystems**, Am Soc Microbiol, v. 4, n. 2, 2019. Disponível em: <<https://msystems.asm.org/content/4/2/e00303-18.abstract>>.
- YANG, X. et al. Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method. **Computational and structural biotechnology journal**, Elsevier, v. 18, p. 153–161, 2020. Disponível em: <<https://doi.org/10.1016/j.csbj.2019.12.005>>.
- YAO, Y. et al. An integration of deep learning with feature embedding for protein–protein interaction prediction. **PeerJ**, PeerJ Inc., v. 7, p. e7126, 2019. Disponível em: <<https://doi.org/10.7717/peerj.7126>>.

- ZENG, M. et al. Protein–protein interaction site prediction through combining local and global features with deep neural networks. **Bioinformatics**, Oxford University Press, v. 36, n. 4, p. 1114–1120, 2020. Disponível em: <<https://doi.org/10.1093/bioinformatics/btz699>>.
- ZHANG, Y. et al. Identification of covid-19 infection-related human genes based on a random walk model in a virus–human protein interaction network. **BioMed research international**, Hindawi, v. 2020, 2020. Disponível em: <<https://doi.org/10.1155/2020/4256301>>.
- ZHENG, N. et al. Targeting virus-host protein interactions: Feature extraction and machine learning approaches. **Current drug metabolism**, Bentham Science Publishers, v. 20, n. 3, p. 177–184, 2019. Disponível em: <<https://doi.org/10.2174/1389200219666180829121038>>.
- ZHOU, Y. et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. **Nature communications**, Nature Publishing Group, v. 10, n. 1, p. 1–10, 2019. Disponível em: <<https://doi.org/10.1038/s41467-019-09234-6>>.

Apêndices

APÊNDICE **A**

Guia do Usuário

O GenPPI é um software autônomo disponibilizado na forma de um arquivo executável para ser utilizado através da linha de comando de um terminal Linux ou do prompt de comando do Windows. O arquivo executável do programa está disponível para download no site: <https://genppi.facom.ufu.br/> ou no repositório: <https://github.com/santosardr/genppi>. Depois de baixá-lo é recomendável criar uma pasta para armazenar o arquivo baixado e os arquivos multi-fasta de proteínas dos genomas que serão analisados, assim como exemplifica a Figura 40.

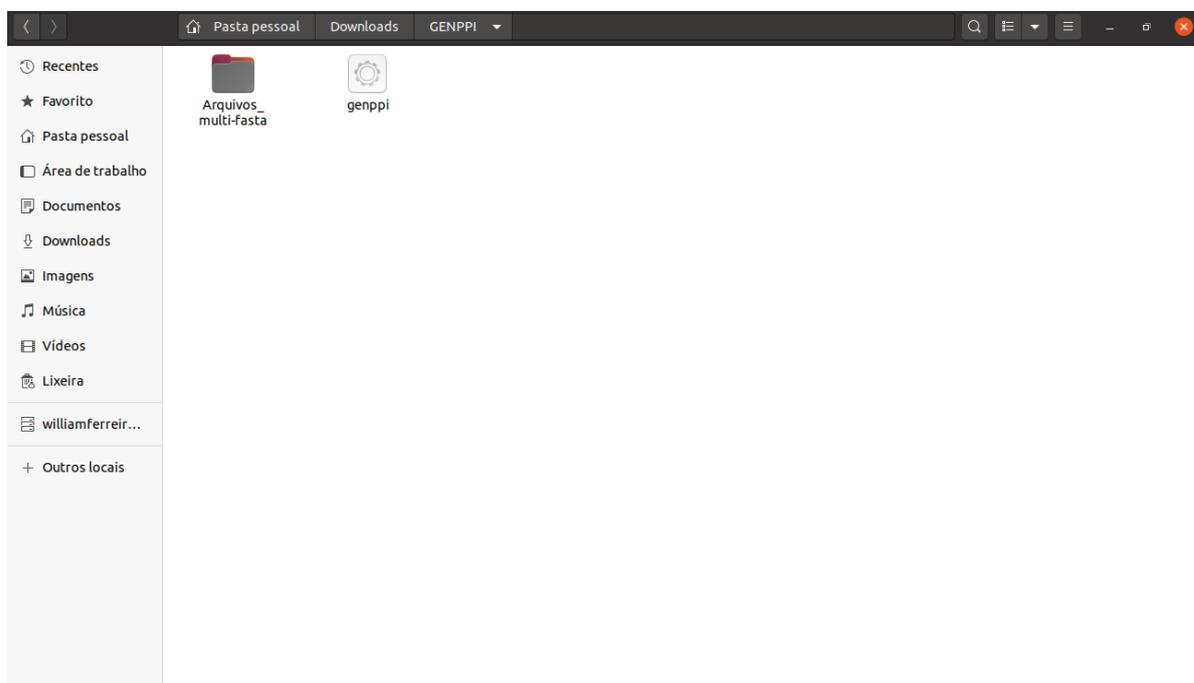


Figura 40 – Criação de um diretório para execução do programa

Os genomas selecionados devem ser representados por arquivos multi-fasta de proteínas, que podem ser baixados em bancos de dados públicos como o NCBI. Tomando como exemplo a Figura 40, todos os arquivos multi-fasta de proteínas devem ser colocados no

diretório Arquivos_multi-fasta para serem lidos pelo programa. Para processar arquivos multi-fasta de proteínas dos genomas selecionados para uma análise, primeiro é necessário acessar o diretório do arquivo executável do programa, através de um terminal Linux ou do prompt de comando do Windows. Estando nesse diretório, ao digitar o nome do executável (genppi) e pressionar enter, é impresso o menu do programa contendo todos os parâmetros que podem ser utilizados, bem como uma explicação para cada um deles (Figura 41).

```
william@william-270ESK: ~/Downloads/GENPPI
william@william-270ESK:~/Downloads/GENPPI$ ./genppi
Mandatory parameter
Directory parameter:
-dir <workingdir> directory
Optional parameters
Parameters of ppi by conserved neighbourhood:
-cnp <conserved neighbourhood score percentage> (65-default)
-expt <type of expansion of the conserved gene neighborhood> 'fixed' or 'dynamic' (fixed-default)
Parameters of conserved neighbourhood by fixed expansion:
-w1 <window size 1> (10-default)
-cw1 <gene conservation required for window 1> (4-default)
-w2 <window size 2> (7-default)
-cw2 <gene conservation required for window 2> (3-default)
-w3 <window size 3> (5-default)
-cw3 <gene conservation required for window 3> (2-default)
-w4 <window size 4> (3-default)
-cw4 <gene conservation required for window 4> (1-default)
Parameters of conserved neighbourhood by dynamic expansion:
-ws <dynamic expansion window size> (1-default)
Parameters of ppi by phylogenetic profile:
-ppp <phylogenetic profiles score percentage> (30-default)
-ppdiff <difference in phylogenetic profiles tolerated to infer ppi> (0-default)
Amino acid histogram parameter settings for the phylogenetic profile:
-pphistofilter <build the phylogenetic profile of genes with a higher percentage of identity>
-ppaadiff <amino acid difference limit> (0-default)
-ppaacheck <minimum amount of amino acids to check> (26-default)
```

Figura 41 – Parâmetros para utilização do GenPPI

Além dos parâmetros mostrados na Figura 41, a outros na sequência que não foram mostrados. Uma explicação detalhada para os parâmetros disponíveis, é dada na Subseção A.1. A diversidade de parâmetros disponíveis tem a função de tornar a predição de redes de PPI mais flexível para demandas de pesquisas específicas do usuário. Por exemplo, o estudo de medidas de centralidade (requer redes com menores densidades) ou o estudo de clusters de proteínas (redes com maiores densidades são mais adequadas). Para executar o programa basta chamar o seu executável na linha de comando seguido da configuração de parâmetros desejável, juntamente com o diretório dos arquivos multi-fasta de proteína através do parâmetro -dir. A Figura 42 demonstra uma execução do programa.

Ao término de uma execução são criados alguns diretórios contendo os resultados das análises do programa como mostra a Figura 43.

Na Figura 43 os arquivos com a extensão .faa são os arquivos multi-fasta de proteínas dos genomas incluídos na análise. Os diretórios existentes foram criados automaticamente em decorrência da execução do programa para esses arquivos multi-fasta. O diretório ppi-files contém redes de interação geradas para todos os genomas (arquivo multi-fasta de proteínas) do conjunto de análise. As redes de interação geradas são disponibilizadas

```

william@william-270E5K: ~/Downloads/GENPPI
william@william-270E5K:~/Downloads/GENPPI$ ./genppi -expt 'dynamic' -ppcomplete -genefusion -dir Arquivos_multi-fasta/

GENPPI VERSION: 1.0
RELEASE NUMBER: 4354980a4dfaede2695e02e525914cead0f2c9ee
REPOSITORY: https://github.com/santosardr/genppi

Directory parameter:
-dir <workingdir> = Arquivos_multi-fasta/

Parameters of ppi by conserved neighbourhood:
-cnp <conserved neighbourhood score percentage> = 65% (default)
-expt <type of expansion of the conserved gene neighborhood> = dynamic
-ws <dynamic expansion window size> = 1 (default)

Parameters of ppi by phylogenetic profile:
Method 2 - PPI prediction by phylogenetic profile without filters
-ppp <phylogenetic profiles score percentage> = 30% (default)
-ppdifftolerated <difference in phylogenetic profiles tolerated to infer ppi> = 0 (default)

Parameters of ppi by gene fusion:
-gfp <gene fusion score percentage> = 5% (default)

Amino acid histogram parameters:
-aadiffllimit <amino acid difference limit> 1 (default)
-aacheckminlimit <minimum amount of amino acids to check> 25 (default)

Making ppi prediction, please wait.

Generating amino acids histogram:
[05 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100]%
[=====]

Generating pan-genome:
[05 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100]%
[=====]

```

Figura 42 – Exemplo de uma execução do GenPPI

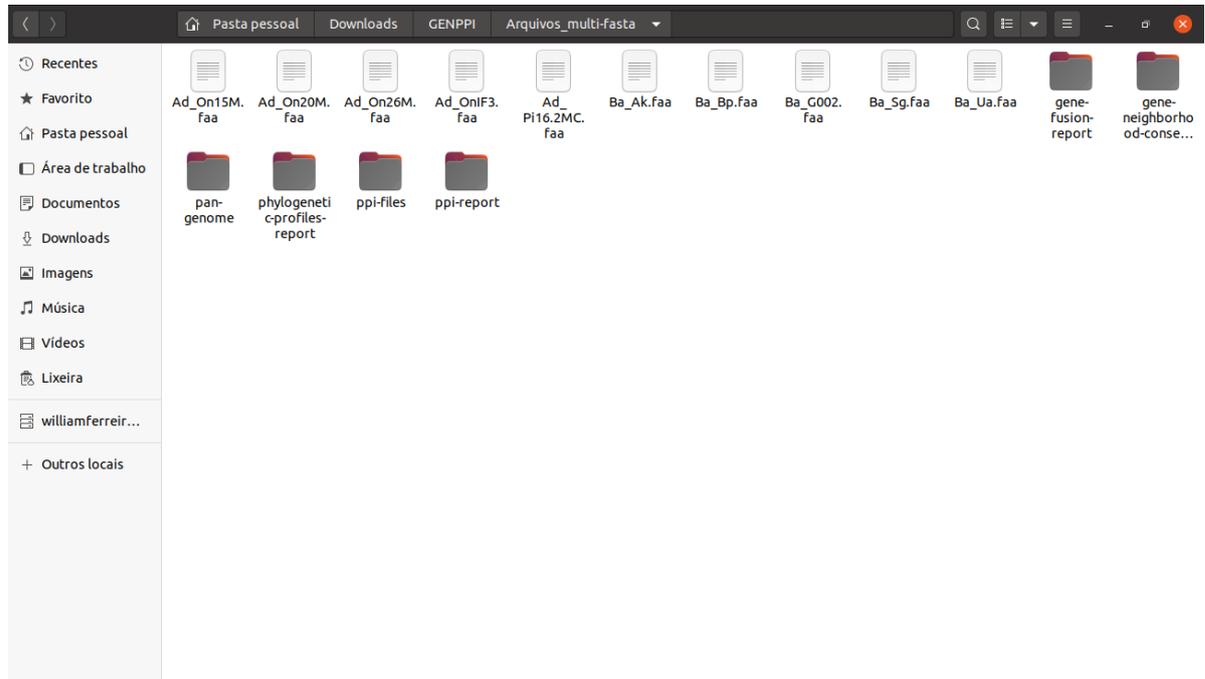


Figura 43 – Diretórios criados em decorrência de uma execução do programa

em arquivos com a extensão `.dot`, que podem ser manipulados pelo programa GEPHI (software para análise de topologia de redes e grafos). Além do diretório com as redes de PPI geradas para os genomas analisados, também são criados outros diretórios contendo diversos tipos de relatórios. Esses contêm informações sobre as interações proteicas inferidas pelos métodos implementados no programa, a saber, os métodos de vizinhança gênica conservada, fusão de genes e perfil filogenético conservado. Também são gerados relatórios sobre o pan-genoma das espécies analisadas. Alguns desses relatórios são a base para a geração dos gráficos utilizados na Subseção 4.2.3 para realizar análises de diferenciação de espécies bacterianas. O usuário também pode gerar esses gráficos para fazer tais análises, através de uma plataforma web criada para esse fim. Basta acessar o site, fazer o upload dos relatórios gerados em uma execução do programa, e utilizar o botão `Submit Query` do site para gerar os gráficos. O site para geração dos gráficos está disponível para acesso público em: <https://genppi.facom.ufu.br/>.

Para que o usuário gere redes de interação utilizando o nosso software, primeiramente é necessário a seleção de um conjunto de genomas de referência para embasar as predições. É importante salientar que a chave para uma aplicação bem-sucedida do GenPPI, é a escolha de um conjunto adequado de genomas de referência, que não devem ser evolutivamente muito distantes, mas não muito próximos uns dos outros. Assim, apenas grupos de genes funcionalmente associados (proteínas interagentes) apresentarão conservação gênica nos genomas de referência e serão inferidos como PPIs.

O restante deste guia segue assim: a Seção A.1 lista e explica todos os parâmetros do GenPPI; e a Seção A.2 apresenta os tipos de relatórios gerados automaticamente após uma execução do programa.

A.1 Parâmetros do GenPPI

Na sequência são apresentados todos os parâmetros do programa e seus significados.

A.1.1 Parâmetros obrigatórios

- ❑ `-dir`: este parâmetro é usado para o usuário informar o diretório dos arquivos multi-fasta de proteínas dos genomas que se deseja analisar. Caso o usuário já esteja dentro desse diretório, basta informar o parâmetro `-dir` sem conteúdo, que o programa usará o diretório atual.

A.1.2 Parâmetros opcionais

Parâmetros para a configuração da heurística proposta:

- ❑ `-aadiffimit`: limite tolerado na diferença de histogramas de aminoácidos.

- `-aacheckminlimit`: quantidade de aminoácidos com diferença de valores de histograma dentro do limite tolerado.

O fator importante na configuração desses dois parâmetros, é o percentual de identidade mínimo garantido por uma possível configuração de valores. As possíveis configurações dos parâmetros da heurística, bem como os percentuais de identidade mínimos garantidos na classificação de similaridade para pares de proteínas, são apresentados pela Tabela 12. As configurações padrões para os parâmetros `-aadifflimit` e `-aacheckminlimit` são, respectivamente, os valores 1 e 25. Estes garantem um percentual de identidade mínima igual a 92,55% para pares de proteínas classificadas como sendo similares pela heurística.

Tabela 12 – Possíveis configurações de parâmetros para a heurística do GenPPI e seus percentuais de identidade mínima garantidos.

<code>-aadifflimit</code> :	<code>-aacheckminlimit</code>	Percentual de identidade mínimo
0	26	100%
0	25	100%
0	24	97,96%
0	23	96,94%
0	22	96,94%
0	21	94,68%
0	20	91,75%
0	19	85,57%
0	18	50,00%
1	26	97,87%
1	25	92,55%

Parâmetros de predição de PPI para a métrica de vizinhança gênica conservada:

- `-cnp`: este parâmetro é responsável por receber o percentual de pontuação atribuído às interações preditas pela métrica de vizinhança gênica conservada. O usuário deve informar o percentual como exemplificado a seguir: `-cnp 50%`. Caso o usuário não informe o parâmetro `-cnp`, o mesmo assumirá seu valor padrão (65%).
- `-expt`: este parâmetro é referente ao tipo de expansão da vizinhança de um gene recorrente. O tipo de expansão pode ser fixo ou dinâmico. O parâmetro `-expt` deve ser informado seguido do tipo de expansão entre aspas simples (`-expt 'fixed'` ou `'dynamic'`). Caso o usuário não informe o parâmetro `-expt`, o mesmo assumirá sua configuração padrão (`'fixed'`).

A expansão fixa dispõe de 4 configurações de janelas diferentes. O objetivo é definir exigências de conservações gênicas distintas, para inferir interações entre proteínas. Os parâmetros disponíveis para configuração da expansão fixa, são:

- ❑ -w1: tamanho da janela 1 (janela principal). A janela 1 é a principal pois é a primeira a ser verificada. O tamanho padrão dessa janela é igual a 10 significando que o algoritmo irá verificar 10 genes vizinhos subsequentes de cada gene recorrente, visando inferir PPIs para os genes que forem achados conservados. O valor padrão do parâmetro -w1 é igual a 10, porém o mesmo pode assumir qualquer valor maior que zero informado pelo usuário. O parâmetro de tamanho da janela -w1 sempre é usado em combinação com o parâmetro -cw1. Este último se refere à quantidade requerida de genes conservados dentro da janela 1 (-w1), para se inferir PPIs para os genes da extensão da janela. O parâmetro -cw1 possui valor de configuração padrão igual a 4 significando uma exigência de 4 genes conservados dentro da janela 1, para se inferir PPIs entre todos os genes da janela. Assim como -w1, o parâmetro -cw1 pode assumir qualquer valor maior que 0, porém deve ser menor ou igual ao tamanho da janela 1.
- ❑ -w2: caso a primeira configuração de janela não seja satisfeita (caso não haja 4 ou mais genes conservados em um intervalo de 10 genes vizinhos subsequentes dos genes recorrentes), então o programa testará a segunda configuração de janela, isto é, os parâmetros -w2 e -cw2. Considerando os valores de configuração padrão para -w2 e -cw2, o algoritmo irá verificar se em um intervalo de 7 genes, existe pelo menos 3 genes conservados junto com o gene recorrente em outros genomas. Caso exista, o programa infere PPIs entre os 7 genes da janela 2. Caso não exista, então o programa testa a configuração da janela 3: os parâmetros -w3 e -cw3.
- ❑ -w3: Considerando os valores de configuração padrão para a janela 3, o programa irá verificar se em um intervalo de 5 genes vizinhos subsequentes de um gene recorrente, existe pelo menos 2 genes conservados. Caso exista, o programa infere PPIs entre os 5 genes da janela 3.
- ❑ -w4: finalmente, caso nenhuma das três primeiras configurações de janela forem satisfeitas, então o programa testa a quarta e última configuração. De acordo com os valores padrões para a configuração da janela 4, o programa irá verificar se em um intervalo de 3 genes, existe pelo menos 1 conservado. Caso exista, o programa infere PPIs entre os 3 genes da janela 4. Caso não exista, então o programa não infere nenhuma PPI para o gene recorrente em questão, e encerra a janela de expansão fixa para esse gene.

Diferentemente da expansão fixa que define tamanhos fixos de janelas para verificar a vizinhança gênica de um gene recorrente, a expansão dinâmica não se limita a um intervalo fixo de genes, continua expandindo a janela progressivamente até que a qualidade de conservação diminua. Para tanto, foi implementado o seguinte parâmetro:

- ❑ -ws: este parâmetro define a quantidade tolerada de genes subsequentes não conservados, para que a expansão dinâmica continue expandindo a janela de verificação da vizinhança de um gene recorrente. O valor de configuração padrão do parâmetro -ws é igual a 3 significando que ao verificar 3 genes vizinhos subsequentes sem identificar nenhum novo gene conservado, a expansão dinâmica deve parar. Porém, caso seja identificado um gene conservado nesse intervalo, então a contagem é reiniciada nesse ponto e verifica-se mais 3 genes subsequentes, e assim por diante. Uma ilustração desse processo foi dada pela Figura 22. Vale lembrar que o valor 3 é a configuração padrão, porém o parâmetro -ws pode assumir qualquer valor maior que zero informado pelo usuário.

Parâmetros de predição de PPI para a métrica de fusão gênica:

- ❑ -genefusion: uma vez que eventos de fusão gênica são mais raros e geram poucas interações proteicas, além do alto custo computacional desse método que pode dobrar o tempo de processamento, a configuração padrão definida não abrange predições pela métrica de fusão gênica. Caso o usuário queira incluir análises de fusão gênica, o parâmetro -genefusion deve ser informado no momento da execução do programa.
- ❑ -gfp: este parâmetro é responsável por receber o percentual de pontuação atribuído às interações preditas pela métrica de fusão gênica. Caso o usuário não informe esse parâmetro, o mesmo assumirá seu valor padrão (5%).

Parâmetros de predição de PPI para a métrica de perfil filogenético conservado:

- ❑ -ppp: este parâmetro é responsável por receber o percentual de pontuação atribuído às interações preditas pela métrica de perfil filogenético conservado. Caso o usuário não informe o parâmetro -ppp, o mesmo assumirá seu valor padrão (30%).
- ❑ -pphistofilter: o uso deste parâmetro é indicado para montar perfis filogenéticos mais confiáveis para os genes, pois o mesmo garante percentuais de identidade mínimos maiores que o padrão estabelecido no programa (>92,55%). Assim, é importante salientar que o seu uso é sempre indicado, pois esse parâmetro produz redes de PPI mais confiáveis. Ele pode ser combinado com os parâmetros -ppaadiffimit e -ppaachekminlimit que definem um percentual de identidade mínima na montagem dos perfis filogenéticos. Porém, o mais indicado é apenas o uso do parâmetro -pphistofilter omitindo os parâmetros -ppaadiffimit e -ppaachekminlimit. Desse modo, esses dois últimos assumirão suas configurações padrões, os valores 0 e 26, respectivamente, garantindo 100% de identidade de sequência na montagem de perfis filogenéticos. Porém, caso o usuário queira relaxar esses dois parâmetros visando prever um número maior de PPIs, os mesmos podem assumir qualquer uma das configurações de valores presentes na Tabela 12.

- -ppdiff tolerated: é sabido que o método de perfil filogenético conservado infere interações para pares de proteínas com perfis filogenéticos idênticos. Porém, pares de proteínas com perfis semelhantes, também podem ser consideradas como interagentes. Para considerar perfis semelhantes nesse método, foi implementado o parâmetro -ppdiff tolerated. Este estabelece a diferença tolerada entre os perfis filogenéticos de um par de proteínas, para classificá-las como interagentes ou não. Um exemplo que ilustra o uso desse parâmetro, pode ser visto na Figura 26 da Subseção 3.7.3.

Para a predição de PPI por meio da métrica de perfil filogenético conservado, foram implementados 7 métodos. Vale destacar que o programa executa utilizando apenas um método por vez, ou seja, o usuário deve escolher o método que melhor atende seu objetivo. Os parâmetros referentes a esses 7 métodos, estão descritos a seguir:

1. -ppcomplete: este parâmetro deve ser utilizado quando se deseja realizar predições de PPIs pelo método de perfil filogenético conservado, sem aplicação de filtros que reduzem o número de interações preditas. Tais aplicações de filtros são feitas de diversas formas através dos parâmetros descritos nos itens a seguir (2 até 7).
2. -ppcn: ao utilizar este parâmetro, serão feitas inferências de PPIs pela métrica de perfil filogenético conservado, apenas para os pares de proteínas já preditas como interagentes pela métrica de vizinhança gênica conservada. Esse parâmetro atua como uma das alternativas de reduzir o número de arestas de PPI no grafo final.
3. -ppiterlimit: este parâmetro estabelece um limite máximo na quantidade de interações preditas para cada genoma. O valor padrão no programa é igual a 500.000 interações, porém qualquer valor maior que zero pode ser informado pelo usuário.
4. -trim: este parâmetro estabelece um limite máximo na quantidade de interações por pontuação estimada, de modo que o número de pares de proteínas interagentes com uma mesma pontuação, não ultrapasse a quantidade definida através desse parâmetro. O valor da configuração padrão no programa para o parâmetro -trim, é igual a 45.000 interações por pontuação, mas pode ser utilizado qualquer valor acima de zero.
5. -threshold: este parâmetro é utilizado para limitar a predição apenas para genes que estejam conservados em um número maior ou menor que um limiar de genomas informado pelo usuário. Por exemplo, se o usuário quiser fazer predições apenas para os genes que estão conservados em mais de 16 genomas, visando omitir interações oriundas de 16 genomas evolutivamente muito mais próximos que os outros do conjunto analisado, então o usuário deve usar o parâmetro -threshold seguido do parâmetro -plusminus '>' com o valor 16. Se for o inverso, ou seja, se o usuário quiser fazer predições apenas para os genes que estão conservados em menos de 16

genomas, então no parâmetro `-plusminus` deve conter um sinal de menor entre aspas simples (`-plusminus '<'`).

6. `-grouplimit`: limite de interações toleradas para manter um grupo de arestas com a mesma pontuação. Esse parâmetro exclui grupos de arestas com um mesmo peso (pontuação), caso suas quantidades totais excederem a um limite estabelecido. Por exemplo, caso o usuário não queira que haja mais de 50 mil arestas de PPIs com uma mesma pontuação na rede de interação, pode-se informar esse limite através do parâmetro `-grouplimit`. Assim, caso haja mais de 50 mil arestas com uma mesma pontuação no grafo final, o programa exclui esse grupo inteiro. O valor da configuração padrão desse parâmetro, é igual a 45.000, podendo o mesmo assumir qualquer valor maior que zero informado pelo usuário.
7. `-profiles`: este parâmetro exclui genes com perfis filogenéticos específicos, caso necessário. Por exemplo, supondo que haja 2 genomas evolutivamente muito próximos no conjunto analisado, isso vai resultar em um número exacerbado de interações para as proteínas desses 2 genomas. Dentre essas muitas interações, provavelmente haverá muitos falsos positivos, pois serão interações resultantes apenas da conservação gênica desses 2 genomas mais próximos, e não resultantes da conservação gênica de todo o conjunto de genomas de referência analisado, como seria o ideal. Para evitar esse problema, o usuário pode usar o parâmetro `-profiles` para descartar perfis filogenéticos específicos na análise, como por exemplo, perfis de genes conservados em apenas 2 genomas. Para tanto, basta usar o parâmetro `-profiles` com o valor 2 (`-profiles 2`), que o programa descartará interações oriundas de perfis filogenéticos conservados em apenas 2 genomas. Caso seja necessário descartar mais de um perfil filogenético, os valores desse parâmetro devem ser colocados entre aspas simples ou duplas e separados por ponto e vírgula: `-profiles "7; 15; 21"`. Neste exemplo serão excluídos genes que estão conservados em 7, 15 e 21 genomas do conjunto analisado.

A.2 Relatórios Gerados Pelo GenPPI

O programa desenvolvido gera alguns relatórios com informações interessantes, como por exemplo: o número de interações preditas por vizinhança gênica conservada, fusão gênica e perfil filogenético; informações sobre o pan-genoma das espécies analisadas, como as proteínas similares dos genomas estudados; informações sobre a conservação da vizinhança gênica de genes recorrentes; informações sobre a fusão de genes; e sobre os perfis filogenéticos das proteínas conservadas encontradas nos genomas. Como a quantidade de dados nesses relatórios pode ser muito grande, na sequência são apresentados apenas trechos de cada relatório visando demonstrá-los. Os dados apresentados são resultantes do processamento de 50 genomas do gênero *Corynebacterium pseudotuberculosis*, uma das bases de dados utilizadas para estudo de caso neste trabalho.

A.2.1 Relatório Sobre o Número de PPIs Preditas Pelos Métodos de Vizinhança Gênica Conservada, Perfil Filogenético, e Fusão Gênica

```

-----Amount of predicted interactions-----
-----
Genome name:          CpFt2193-67-STRING
Number of ppi by cn: 10368
Number of ppi by pp: 111187
Number of ppi by gf: 4
-----
Genome name:          GCA_000006605.1_ASM660v1_protein
Number of ppi by cn: 265
Number of ppi by pp: 75
Number of ppi by gf: 0
-----
Genome name:          GCA_000011305.1_ASM1130v1_protein
Number of ppi by cn: 300
Number of ppi by pp: 1
Number of ppi by gf: 1
-----
Genome name:          GCA_000011325.1_ASM1132v1_protein
Number of ppi by cn: 1065
Number of ppi by pp: 101
Number of ppi by gf: 0
-----
Genome name:          GCA_000022905.1_ASM2290v1_protein
Number of ppi by cn: 294
Number of ppi by pp: 4
Number of ppi by gf: 0
-----
Genome name:          GCA_000023145.1_ASM2314v1_protein
Number of ppi by cn: 46
Number of ppi by pp: 0
Number of ppi by gf: 0
-----

```

Figura 44 – Trecho do relatório sobre o número de PPIs preditas pelas métricas de vizinhança gênica conservada, perfil filogenético e fusão gênica, para genomas de um conjunto sob análise

A.2.2 Relatório Sobre o Pan-Genoma

```

-----Pan genome-----
Minimum identity percentage: 92,55%
-----
Protein: CPTC_00001 | Genome: CpFt2193-67-STRING
Similar proteins:
Protein: AEP70879.1 | Genome: GCA_000227175.1_ASM22717v1_protein;
Protein: AEP69659.1 | Genome: GCA_000233735.1_ASM23373v1_protein;
Protein: AEX40144.1 | Genome: GCA_000241855.1_ASM24185v1_protein;
-----
Protein: CPTC_00002 | Genome: CpFt2193-67-STRING
Similar proteins:
Protein: ADK29442.1 | Genome: GCA_000143705.2_ASM14370v2_protein;
Protein: ADL21515.1 | Genome: GCA_000144935.3_ASM14493v3_protein;
Protein: AEP70880.1 | Genome: GCA_000227175.1_ASM22717v1_protein;
Protein: AEX40145.1 | Genome: GCA_000241855.1_ASM24185v1_protein;
Protein: AFH52597.1 | Genome: GCA_000258385.1_ASM25838v1_protein;
-----

```

Figura 45 – Trecho do relatório sobre o pan-genoma de espécies incluídas em uma análise

A.2.3 Relatório Sobre Vizinhança Gênica Conservada

```

#-----Gene neighborhood conservation report-----
#Minimum identity percentage: 92,55%, Expansion type: fixed, Window size: 10
#-----
#Columns per row: Genome, Gene, Number of genes conserved, Number of genes not conserved, Number of genomes of each conserved gene;
CpFt2193-67-STRING, CPTC_00001, 9, 1, 3 4 2 2 2 3 4 4 4;
CpFt2193-67-STRING, CPTC_00002, 8, 2, 6 2 2 5 5 5 5 4;
CpFt2193-67-STRING, CPTC_00003, 9, 1, 2 2 2 7 8 8 8 7 7;
CpFt2193-67-STRING, CPTC_00005, 8, 2, 2 2 2 2 2 2 2 2;
CpFt2193-67-STRING, CPTC_00006, 8, 2, 2 2 2 2 2 2 2 2;
CpFt2193-67-STRING, CPTC_00007, 7, 3, 2 2 2 2 2 2 2;
CpFt2193-67-STRING, CPTC_00008, 10, 0, 13 13 13 10 12 2 13 6 2 11;
CpFt2193-67-STRING, CPTC_00009, 10, 0, 16 16 13 15 2 17 7 2 13 16;
CpFt2193-67-STRING, CPTC_00010, 10, 0, 16 13 15 2 16 7 2 13 16 2;
CpFt2193-67-STRING, CPTC_00011, 9, 1, 13 15 2 16 7 2 13 16 2;
CpFt2193-67-STRING, CPTC_00012, 9, 1, 12 2 13 4 2 11 13 2 3;
CpFt2193-67-STRING, CPTC_00013, 6, 4, 15 6 12 15 5 3;
CpFt2193-67-STRING, CPTC_00014, 9, 1, 2 2 2 2 2 2 2 2 2;
CpFt2193-67-STRING, CPTC_00015, 9, 1, 7 2 13 16 2 6 4 5 6;
CpFt2193-67-STRING, CPTC_00016, 9, 1, 2 4 7 2 6 4 5 6 6;
CpFt2193-67-STRING, CPTC_00017, 9, 1, 2 2 2 2 2 2 2 2 2;

```

Figura 46 – Trecho do relatório sobre a conservação da vizinhança gênica de genes recorrentes, em uma janela de expansão fixa de tamanho 10

A.2.4 Relatório Sobre Interações Preditas Pelo Método de Fusão Gênica

```

-----Gene fusion report-----
Minimum identity percentage: 92,55%
-----
Genome:      CpFt2193-67-STRING
Fusion 1 =>  PPI: CPTC_00468 -- CPTC_00469
              Rosetta-stone: ADK27987.1 in GCA_000143705.2_ASM14370v2_protein
              Rosetta-stone: ADL20099.1 in GCA_000144935.3_ASM14493v3_protein
              Rosetta-stone: AEX38672.1 in GCA_000241855.1_ASM24185v1_protein
              Rosetta-stone: AFH51114.1 in GCA_000258385.1_ASM25838v1_protein
Fusion 2 =>  PPI: CPTC_00733 -- CPTC_00734
              Rosetta-stone: ADK29850.1 in GCA_000143705.2_ASM14370v2_protein
              Rosetta-stone: ADL11499.2 in GCA_000144675.2_ASM14467v2_protein
              Rosetta-stone: ADL21912.2 in GCA_000144935.3_ASM14493v3_protein
              Rosetta-stone: AD027309.3 in GCA_000152065.3_ASM15206v3_protein
              Rosetta-stone: AEK93369.1 in GCA_000221625.1_ASM22162v1_protein
              Rosetta-stone: AEP71275.1 in GCA_000227175.1_ASM22717v1_protein
              Rosetta-stone: AEQ07576.2 in GCA_000227605.3_ASM22760v3_protein
              Rosetta-stone: AEX40563.1 in GCA_000241855.1_ASM24185v1_protein
              Rosetta-stone: AFB73390.2 in GCA_000248375.2_ASM24837v2_protein
              Rosetta-stone: AFF23199.1 in GCA_000255935.1_ASM25593v1_protein
              Rosetta-stone: AFH91850.2 in GCA_000259155.4_ASM25915v4_protein
              Rosetta-stone: AFK17688.2 in GCA_000263755.3_ASM26375v3_protein
              Rosetta-stone: AFM08341.3 in GCA_000265545.3_ASM26554v3_protein

```

Figura 47 – Trecho do relatório sobre interações preditas pelo método de fusão gênica

A.2.5 Relatório Sobre o Perfil Filogenético das Proteínas Conservadas de Cada Genoma Incluso em Uma Análise

```

-----Report of complete ppi prediction by phylogenetic profile-----
Minimum identity percentage: 100%
-----

Genome: GCA_000006605.1_ASM660v1_protein
Total profiles: 8
Total ppi: 75
-----
Profile 1  => Total genes: 1   | Total ppi: 0   | Weight: 0.000 | Number of genomes: 10;
Profile 2  => Total genes: 1   | Total ppi: 0   | Weight: 0.000 | Number of genomes: 3;
Profile 3  => Total genes: 1   | Total ppi: 0   | Weight: 0.000 | Number of genomes: 4;
Profile 4  => Total genes: 1   | Total ppi: 0   | Weight: 0.000 | Number of genomes: 2;
Profile 5  => Total genes: 1   | Total ppi: 0   | Weight: 0.000 | Number of genomes: 5;
Profile 6  => Total genes: 5   | Total ppi: 10  | Weight: 0.200 | Number of genomes: 2;
Profile 7  => Total genes: 5   | Total ppi: 10  | Weight: 0.200 | Number of genomes: 2;
Profile 8  => Total genes: 11  | Total ppi: 55  | Weight: 0.300 | Number of genomes: 4;
-----

Genome: GCA_000011305.1_ASM1130v1_protein
Total profiles: 4
Total ppi: 1
-----
Profile 1  => Total genes: 1   | Total ppi: 0   | Weight: 0.000 | Number of genomes: 2;
Profile 2  => Total genes: 1   | Total ppi: 0   | Weight: 0.000 | Number of genomes: 5;
Profile 3  => Total genes: 1   | Total ppi: 0   | Weight: 0.000 | Number of genomes: 10;
Profile 4  => Total genes: 2   | Total ppi: 1   | Weight: 0.300 | Number of genomes: 2;
-----

```

Figura 48 – Trecho do relatório que fornece informações sobre o perfil filogenético das proteínas conservadas de genomas analisados. Esse relatório é gerado apenas quando se utiliza o parâmetro `-ppcomplete`

Gráficos

Neste trabalho, três estudos de casos foram realizados visando verificar a correção e confiabilidade biológica de dados gerados pelo programa, para inferir redes de PPI. Nesses estudos de caso analisou-se genomas de três gêneros bacterianos, a saber, *Dietzia*, *Corynebacterium* e *Aeromonas*. Na Subseção 4.2.3, utilizou-se três tipos de gráficos para apresentar os resultados das análises computacionais realizadas para os gêneros bacterianos estudados. Além desses gráficos, outros três também foram gerados nessas análises. Esses não foram incluídos no capítulo de experimentos e análise de resultados, simplesmente por refletirem dados já demonstrados pelos outros gráficos, não acrescentando nenhum resultado valioso para validar a hipótese principal deste trabalho.

O objetivo deste apêndice é apresentar os gráficos restantes demonstrando resultados obtidos nos três estudos de caso realizados.

B.1 Gráficos

A seguir, é dada uma descrição dos gráficos que serão apresentados para análise de resultados.

B.1.1 Gráfico de Barras Sobre a Quantidade de Interações Proteicas Inferidas Pela Método de Vizinhança Gênica Conservada

Feito a partir do relatório demonstrado na Subseção A.2.1, esse gráfico representa a soma de todas as interações preditas pelo método de vizinhança gênica conservada. Essas interações são contabilizadas como $n*(n-1)/2$, com n sendo o número de genes distribuídos nos intervalos recorrentes de genes (*operons*). Quanto maior o número de genes conservados nesses intervalos, maior será a quantidade de interações possíveis. Para as predições de PPI pelo método de vizinha gênica conservada, considera-se que as proteínas dos arquivos multi-fasta dos genomas analisados, estejam ordenadas tal qual os seus ge-

nes estão dispostos em uma fita de DNA. Essa consideração é coerente porque proteínas geralmente entram em um arquivo multi-*fasta*, em uma sequência similar a que estavam quando extraídas da fita de DNA. Essa análise está considerando uma janela de expansão fixa de tamanho igual a 7 genes vizinhos subsequentes a um gene recorrente.

B.1.2 Histograma da Quantidade de Perfis Filogenéticos Encontrados nos Genomas

Feito a partir do relatório demonstrado na Subseção A.2.5, aqui temos uma série de gráficos de histograma representando os perfis filogenéticos das proteínas conservadas dos genomas. Esses gráficos são próximos de uma “impressão digital” dos genomas, baseada nos perfis filogenéticos encontrados em cada genoma. Esse gráfico está em escala logarítmica porque existem perfis filogenéticos com muitos genes (centenas) e outros com pouquíssimos genes (basta ser > 1). Aqui o que importa são padrões gráficos que podem ser notados comparando um gráfico com os demais. Gráficos que possuem padrões de barras (histograma) semelhantes são uma evidência de genomas evolutivamente próximos.

B.1.3 Gráfico de Barras Sobre o Total de Interações Proteicas Inferidas pelo Método de Perfil Filogenético Conservado

Feito a partir do relatório demonstrado na Subseção A.2.1, esse gráfico apresenta para um genoma, a somatória das interações inferidas para todos os possíveis pares de proteínas com perfis filogenéticos conservados. Um perfil filogenético encontrado em um dado genoma, é conservado se as proteínas com esse perfil filogenético, estiverem conservadas em dois ou mais genomas do conjunto analisado. A quantidade total de interações inferidas para os genomas, pelo método de perfil filogenético conservado, costuma ser bastante expressiva por ser o resultado da somatória das $n*(n-1)/2$ possíveis interações dos perfis identificados. Sendo n o número de proteínas de cada perfil filogenético conservado encontrado em um genoma. Assim, dois genomas evolutivamente muito próximos, podem gerar perfis filogenéticos conservados com centenas de genes que se transformam em milhares de prováveis interações criando uma notoriedade visual frente aos demais genomas nesse gráfico.

B.2 Resultados Obtidos nos Estudos de Caso Realizados

Nesta seção é apresentada uma análise de resultados dos gráficos descritos na seção anterior, gerados a partir de dados obtidos pelo programa nos três estudos de caso reali-

zados. Assim, na sequência são mostrados resultados obtidos para os gêneros de bactéria estudados.

B.2.1 Estudo de Caso 1 – Gênero Bacteriano *Dietzia*

A quantidade de interações inferidas pelo método de vizinhança gênica conservada, reflete o nível de conservação gênica de um genoma em relação aos demais. Portanto, podemos perceber que no gráfico da Figura 49 os genomas de *Rhodococcus* estão com uma quantidade maior de conservação gênica entre si, pois os genomas dessa espécie apresentam quantidades próximas de interações entre suas proteínas (barras niveladas). Essa conservação gênica indicada para os genomas de *Rhodococcus*, está também evidenciada no gráfico boxplot da Figura 30 apresentado na Subseção 4.2.3.1. Curiosamente a pior das conservações gênicas fica com um genoma de *Rhodococcus*, o primeiro genoma apresentado no gráfico da Figura 49. Isso pode ser um sinal de que essa seja outra espécie ou a montagem depositada no banco de dados público, estava com baixa qualidade.

Em meados de Junho de 2020, o grupo RECOM estava tentando publicar o genoma *Dietzia_sp_123*, um genoma até então inédito. Curiosamente esse genoma está muito próximo aos genomas com as menores quantidades de conservação gênica. Na lista ordenada ele seria o sexto genoma com a menor conservação frente ao grupo de bactérias analisadas. Essa baixa conservação frente ao grupo, pode significar que o genoma ainda está com uma qualidade de montagem inferior ao necessário ou que trata-se de uma espécie muito diferente das demais. O grupo RECOM, até o mês de Junho de 2020, ainda estava tentando esgotar as possibilidades de melhoria de montagem desse genoma para concluir entre essas possibilidades.

Os gráficos da Figura 50 conseguem mostrar particularidades entre espécies de bactérias de ambos os gêneros. Mesmo genomas da mesma espécie possuem perfis muito distintos. Observando os padrões de barras dos gráficos da Figura 50, decorrentes dos perfis filogenéticos encontrados nos genomas, é possível identificar quais são de fato evolutivamente próximos ou distantes. Essas análises podem ser utilizadas como suporte para definição de subespécies. Entretanto, como a definição de subespécies ou cepas bacterianas carece de outras análises (análises bioquímicas), aguardamos uma análise mais aprofundada para levantar a hipótese de existência de diferentes cepas ou espécies dentre os genomas desses dois gêneros (*Dietzia* e *Rhodococcus*).

Quando dois genomas são evolutivamente muito próximos, esses terão um número muito grande de interações por perfil filogenético conservado se destacando dos demais genomas no gráfico da Figura 51. Foi o que aconteceu com quatro genomas de *Dietzia* localizados nas quatro últimas posições desse gráfico. Provavelmente são dois pares de genomas muito similares. Novamente o *Rhodococcus* está com uma quantidade uniforme de interações entre todos os representantes do gênero nesse conjunto de dados. As espécies menos representadas (à esquerda do gráfico) são genomas com perfis filogenéticos que oco-

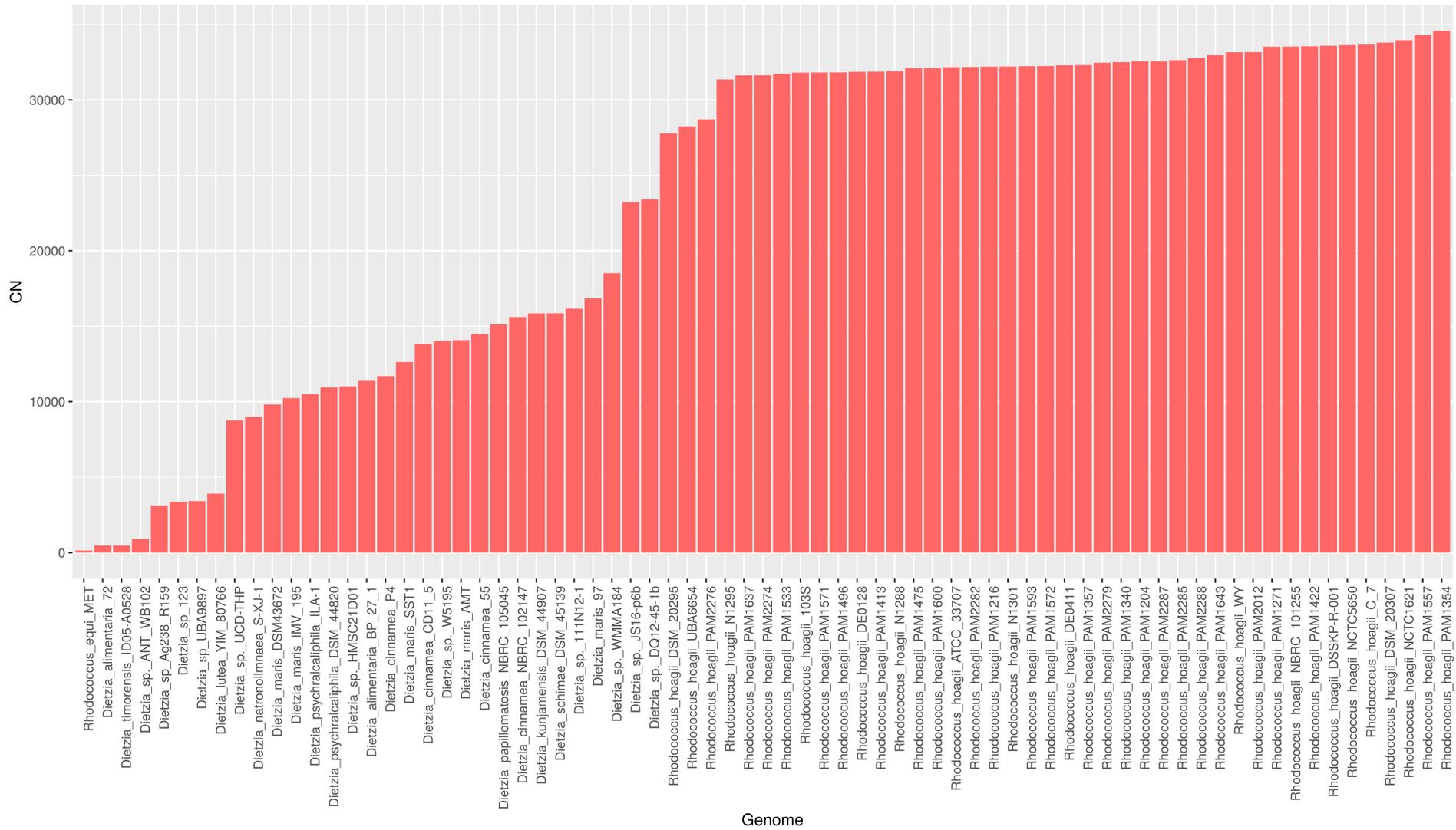


Figura 49 – Gráfico de barras sobre a quantidade de interações proteicas inferidas pelo método de vizinhança gênica conservada.

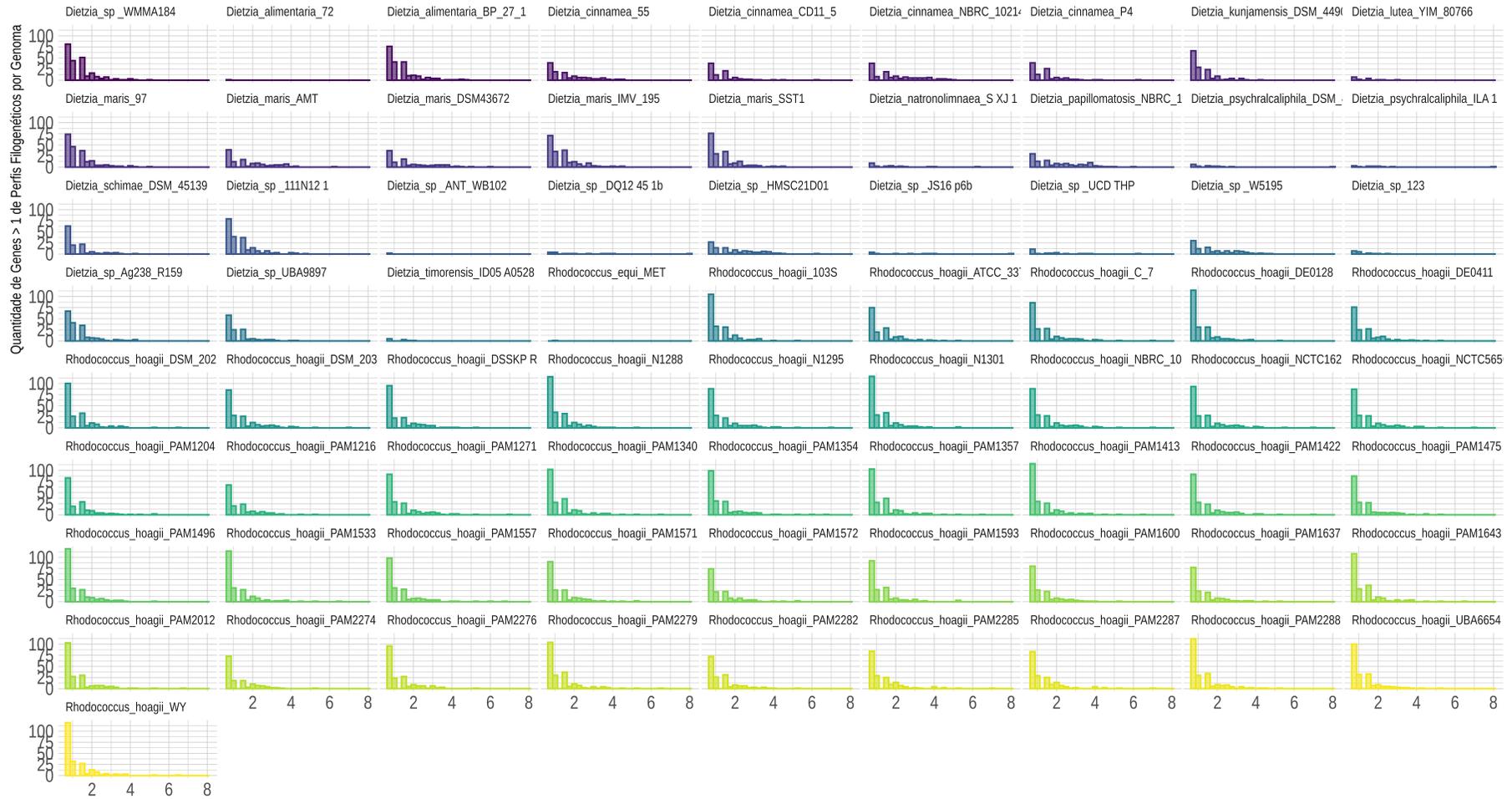


Figura 50 – Histograma da quantidade de perfis filogenéticos encontrados nos genomas

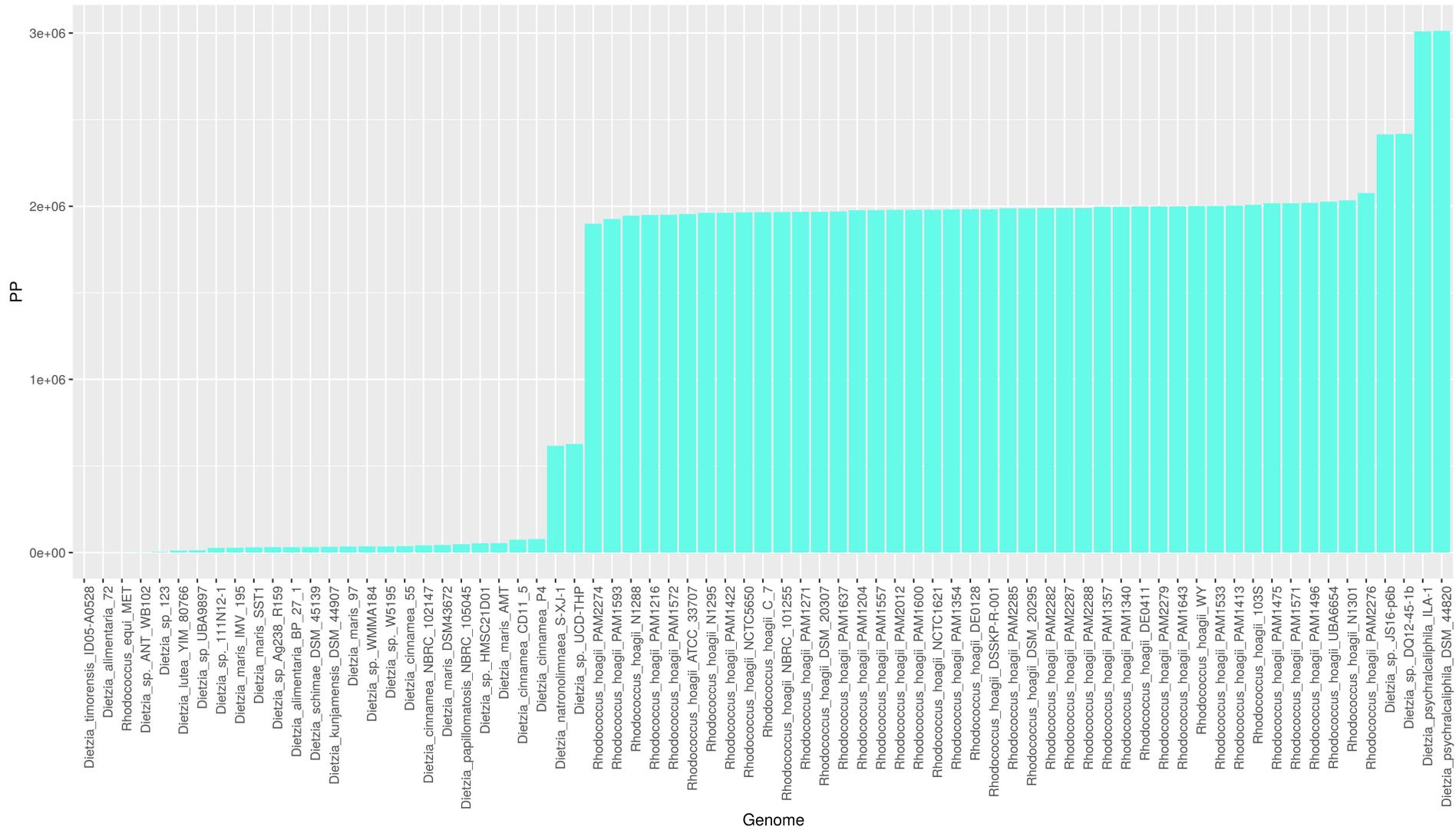


Figura 51 – Gráfico de barras sobre o total de interações proteicas inferidas pelo método de perfil filogenético conservado

rem em baixa frequência nesse conjunto de dados. Esse é o outro indicativo de que seriam espécies ou cepas bastante diferentes das demais ao centro e à direita do gráfico. A título de exemplo cito as espécies *Dietzia timorensis*, *Dietzia alimentaria* e *Rhodococcus equi*. Esses três genomas estão no canto esquerdo do gráfico da Figura 51. Esses mesmos genomas também estão com um padrão de barras muito baixa no gráfico da Figura 50. É evidente que esses três genomas são de espécies distintas das demais desse conjunto de genomas, sendo os únicos representantes de suas espécies.

B.2.2 Estudo de Caso 2 – Gênero Bacteriano *Corynebacterium*

No gráfico da Figura 52, percebe-se uma vizinhança gênica conservada mais significativa entre os genomas de *C. Pseudotuberculosis* (barras de cores laranja (biovar *ovis*) e amarelo (biovar *equi*)), seguido pelos genomas de *C. Diphtheriae* (barras de cor verde). As demais espécies possuem apenas um genoma representando-as nesse conjunto de 50 genomas do gênero *Corynebacterium*. Essas demais espécies, por estarem menos representadas nesse conjunto de genomas, não mostraram conservação de vizinhança gênica expressiva e, portanto, possuem barras menos elevadas.

Através dos gráficos da Figura 53, existe a possibilidade de fazermos distinção entre espécies de bactérias, inclusive entre biovars (subespécies) de uma espécie. Para tanto, basta notar os padrões gráficos dos genomas do grupo de análise. Os genomas de *C. Pseudotuberculosis* são todos os que possuem o prefixo Cp. Nesse gráfico podemos ver uma diferença nos padrões de barras entre os genomas do biovar *ovis* e *equi* de *C. Pseudotuberculosis* (BERNARDES et al., 2020). Os genomas do biovar *ovis* são os que apresentam um padrão de barras mais elevado, sendo esses os seguintes genomas: Cp119, Cp1002B, Cp231, Cp267, Cp3995, Cp4202A, CpFRC41, CpPAT10, CpString e P54B96. Os genomas do biovar *equi* são os de nome Cp106A, CP162, Cp258, Cp31, Cp316 e Cp-CIP5297. Estes últimos apresentam um padrão de barras menos elevado em relação aos genomas do biovar *ovis*. Essa diferenciação entre os genomas do biovar *ovis* e *equi* de *C. Pseudotuberculosis*, corrobora com os resultados dos gráficos das Figuras 32 e 33 da Subseção 4.2.3.2. Nesses dois gráficos também foi possível perceber essa distinção de biovars. Outros padrões gráficos também podem ser notados para os genomas de *C. Diphtheriae*, e para o restante dos genomas de *Corynebacterium* do grupo de análise. Esses possuem, respectivamente, os prefixos Cdip e GCA.

O gráfico da Figura 54 destaca que as linhagens de *C. pseudotuberculosis* do biovar *equi*, com os nomes Cp258 (oriunda de um cavalo nos EUA) e CpCIP5297 (oriunda de um cavalo no Quênia), representadas pelas duas barras amarelas de maior altura, são muito similares entre si, mais similares do que as demais. A alta quantidade de genes conservados nesses dois genomas, resultou em um número proporcional e muito grande de possíveis interações entre os genes de cada um desses dois genomas. Isso poderia sugerir um subtipo dentre as espécies de *C. pseudotuberculosis* do biovar *equi*.

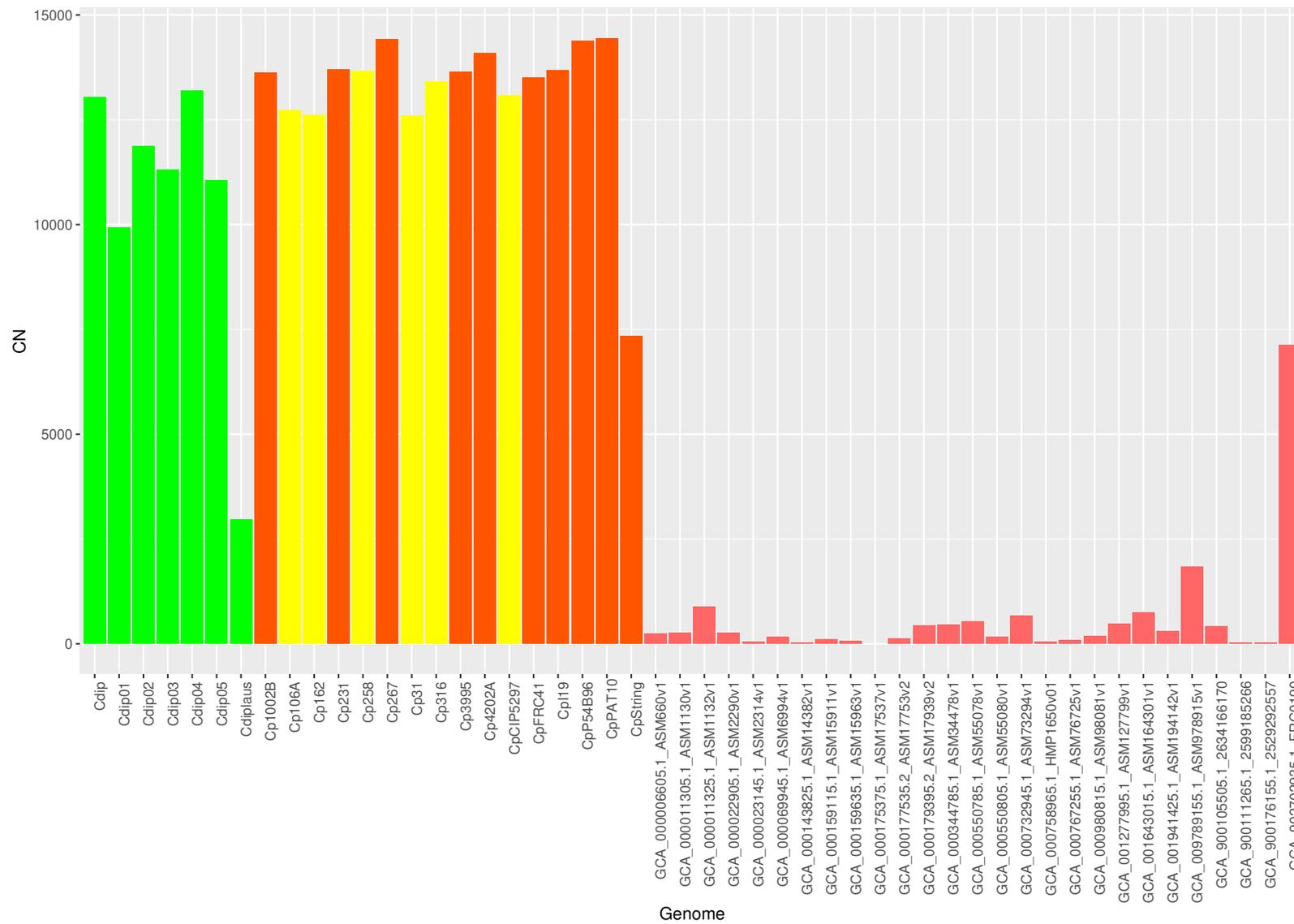


Figura 52 – Gráfico de barras sobre a quantidade de interações proteicas inferidas pelo método de vizinhança gênica conservada.

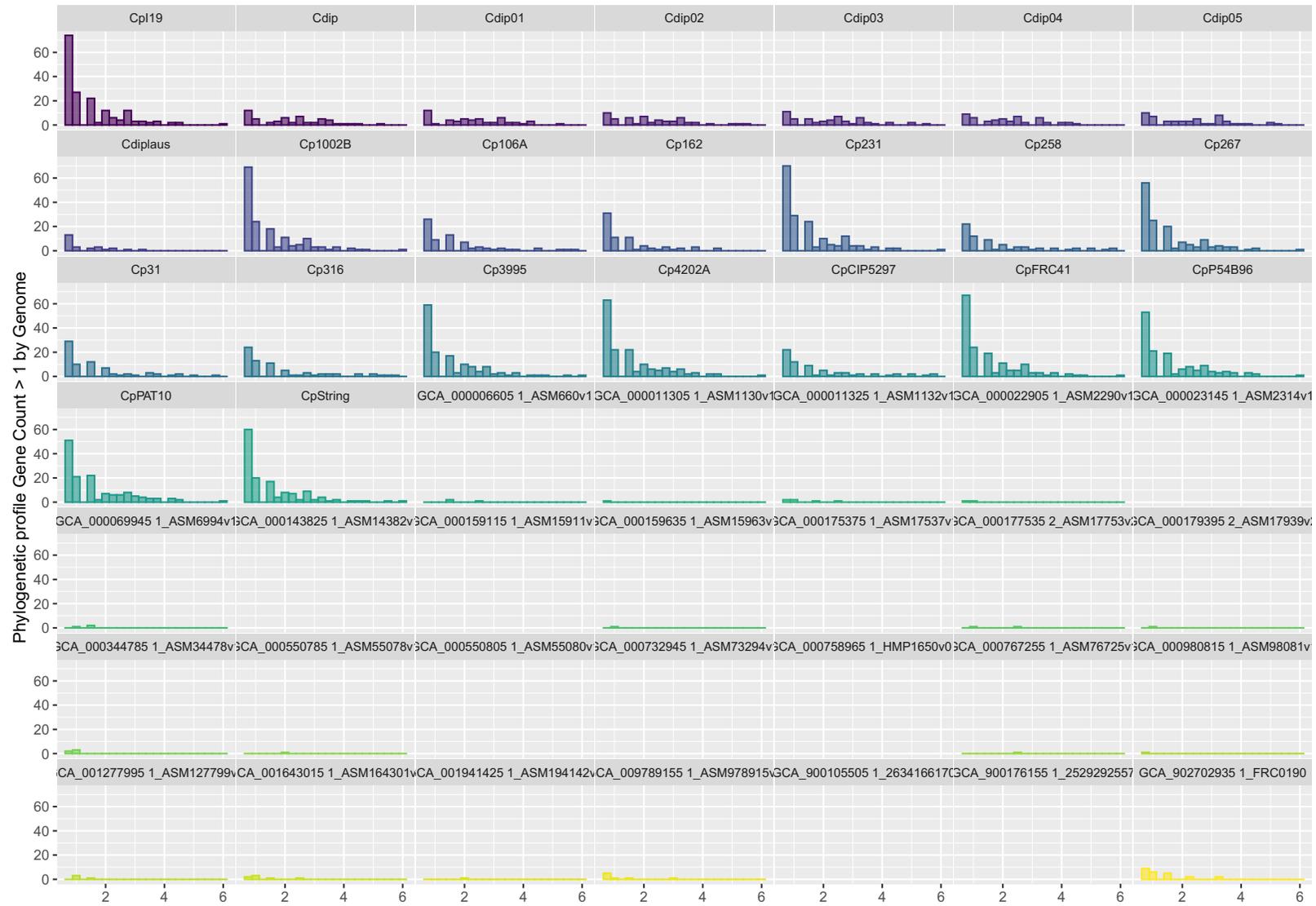


Figura 53 – Histograma da quantidade de perfis filogenéticos encontrados nos genomas

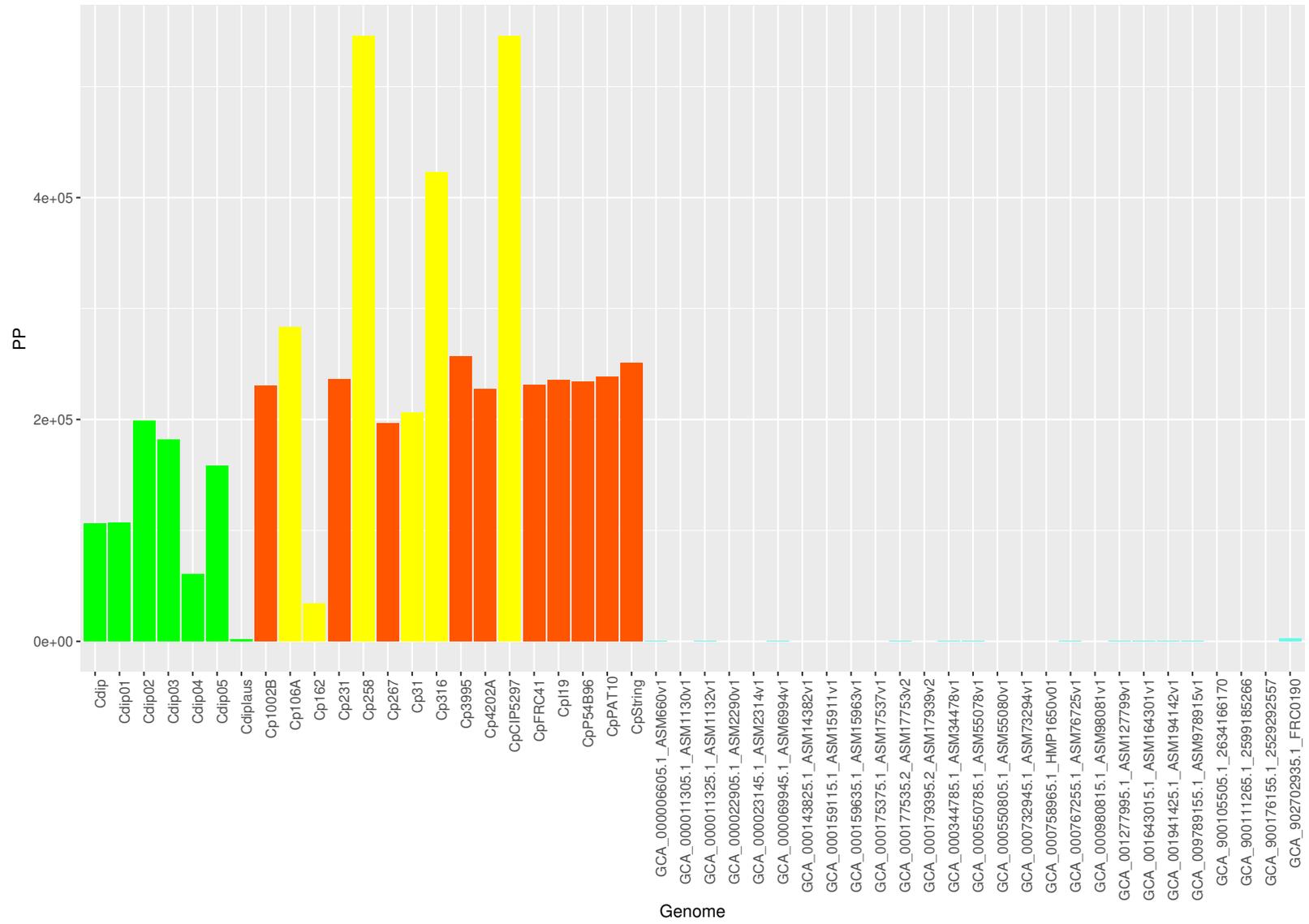


Figura 54 – Gráfico de barras sobre o total de interações proteicas inferidas pelo método de perfil filogenético conservado

B.2.3 Estudo de Caso 3 – Gênero Bacteriano *Aeromonas*

No gráfico da Figura 55 referente ao histograma da quantidade de interações inferidas para cada genoma pelo método de vizinhança gênica conservada, quatro genomas de *A. hydrophila* (barras amarelas) se destoaram do perfil do restante de genomas dessa linhagem. Esses quatro são os seguintes: Ah_On16M, Ah_On17M, Ah_On18M e Ah_Pi13.2HPAS. Esse gráfico evidencia uma menor vizinhança gênica conservada para esses quatro genomas da espécie *A. hydrophila*, em relação ao restante de genomas dessa espécie. O gráfico boxplot da Figura 35 apresentado na Subseção 4.2.3.3, também indicou que esses quatro genomas tiveram uma quantidade menor de perfis filogenéticos conservados em relação aos outros genomas de *A. hydrophila*. Ao que parece, esses 4 genomas podem ser um subtipo de *A. hydrophila* ou até mesmo uma espécie inédita de bactéria. Situação similar acontece com um genoma de *A. veronii* denominado Av_Pi4.2HPAS e representado pela barra de cor verde.

Os gráficos da Figura 56, apresentam um padrão de barras muito uniforme mesmo entre genomas de espécies distintas do gênero *Aeromonas*, com raras exceções.

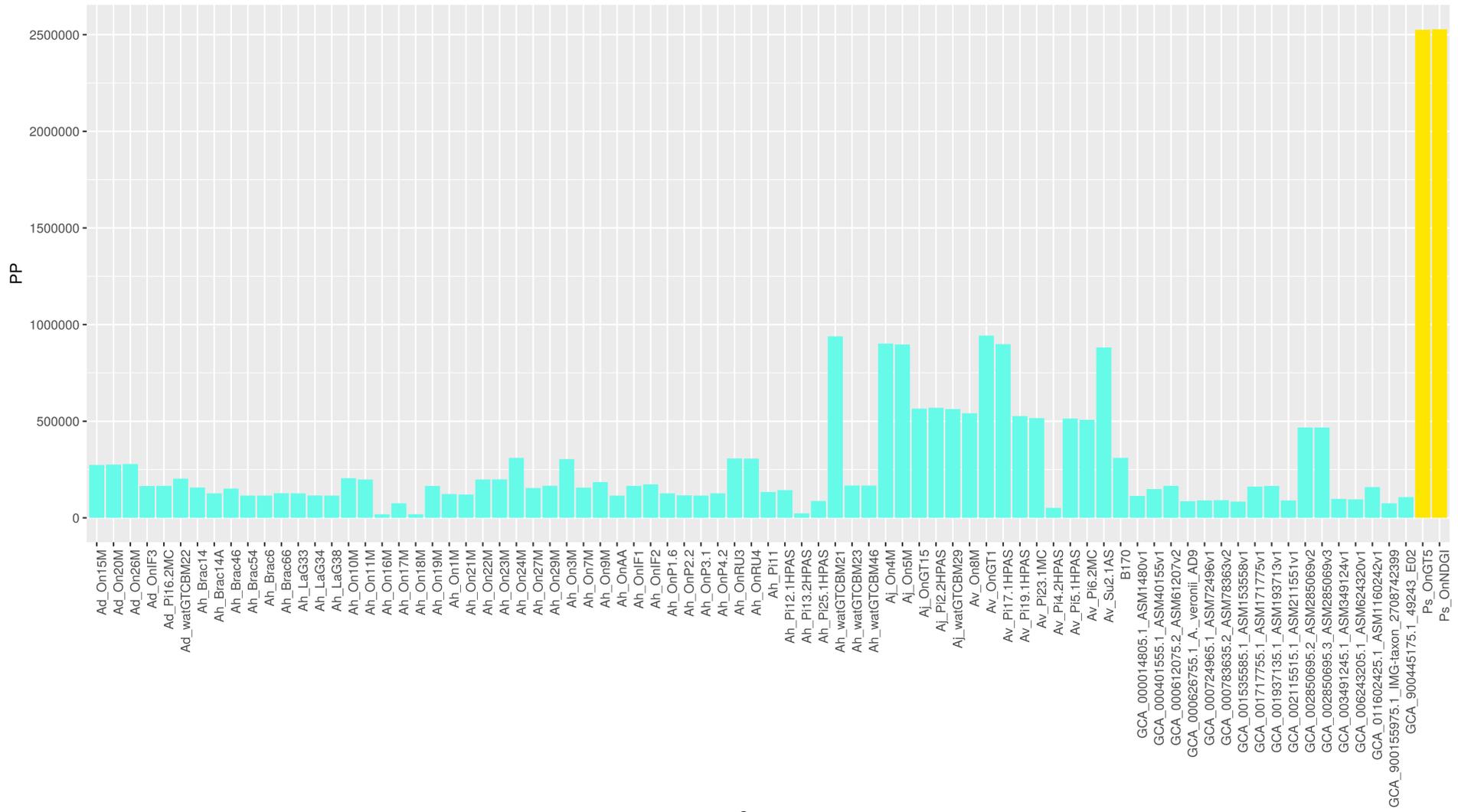
Os dois genomas que se destacam no gráfico da Figura 57 (barras amarelas), são genomas de espécies diferentes do gênero de *Aeromonas*, e evolutivamente próximos. Esses dois foram inseridos no grupo de análise por engano, mas que serviram como mais uma evidência de que as predições feitas pelo GenPPI procedem. Portanto, esses dois genomas inseridos por engado, não foram excluídos do grupo de análise. Como a quantidade de perfis conservados apenas entre esses dois genomas, era muito grande, a quantidade de possíveis interações apenas para esses dois genomas, extrapolou muito a média dos demais genomas. Sempre que há duas barras com alturas muito próximas nesse gráfico, é sinal que a quantidade de perfis filogenéticos conservados entre os dois genomas em questão, é alta evidenciando uma maior proximidade evolutiva para os mesmos.



Figura 55 – Gráfico de barras sobre a quantidade de interações proteicas inferidas pelo método de vizinhança gênica conservada.



Figura 56 – Histograma da quantidade de perfis filogenéticos encontrados nos genomas



Genome

Figura 57 – Gráfico de barras sobre o total de interações proteicas inferidas pelo método de perfil filogenético conservado