UNIVERSIDADE ESTADUAL DE PONTA GROSSA
SETOR DE CIÊNCIAS AGRÁRIAS E DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

LUIZ OTÁVIO OYAMA

DESENVOLVIMENTO DE UMA FERRAMENTA PARA IDENTIFICAR PROTEÍNAS RIBOSSOMAIS EM ESPECTRO DE MASSA DO TIPO MALDI-TOF

PONTA GROSSA

2023

LUIZ OTÁVIO OYAMA

DESENVOLVIMENTO DE UMA FERRAMENTA PARA IDENTIFICAR PROTEÍNAS RIBOSSOMAIS EM ESPECTRO DE MASSA DO TIPO MALDI-TOF

Dissertação apresentada ao Programa de Pós-Graduação em Computação Aplicada, curso de Mestrado em Computação Aplicada da Universidade Estadual de Ponta Grossa, como requisito parcial para obtenção do título de Mestre.

Orientação: Prof. Dr. Rafael Mazer Etto

PONTA GROSSA

2023

Oyama, Luiz Otávio
O98 Desenvolvimento de uma ferramenta para identificar proteínas ribossomais em espectro de massa do tipo MALDI-TOF / Luiz Otávio Oyama. Ponta Grossa, 2023.
53 f.

Dissertação (Mestrado em Computação Aplicada - Área de Concentração: Computação para Tecnologias em Agricultura), Universidade Estadual de Ponta Grossa.

Orientador: Prof. Dr. Rafael Mazer Etto.

1. Aprendizado de máquina. 2. Bioinformática. 3. Biotecnologia. 4. Espectrometria de massa. I. Etto, Rafael Mazer. II. Universidade Estadual de Ponta Grossa. Computação para Tecnologias em Agricultura. III.T.

CDD: 004

Ficha catalográfica elaborada por Maria Luzia Fernandes Bertholino dos Santos- CRB9/986

UNIVERSIDADE ESTADUAL DE PONTA GROSSA
Av. General Carlos Cavalcanti, 4748 - Bairro Uvaranas - CEP 84030-900 - Ponta Grossa - PR - https://uepg.br

TERMO

TERMO DE APROVAÇÃO

Luiz Otávio Oyama

DESENVOLVIMENTO DE UMA FERRAMENTA PARA IDENTIFICAR PROTEÍNAS RIBOSSOMAIS EM ESPECTRO DE MASSA MALDI-TOF

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre no Programa de Pós-Graduação em Computação Aplicada da Universidade Estadual de Ponta Grossa, pela seguinte banca examinadora:

Prof. Dr. Rafael Mazer Etto (UEPG - Presidente)

Prof. Dr. José Carlos Ferreira da Rocha (UEPG)

Prof. Dr. Wellington Claiton Leite (OAK RIDGE NATIONAL LABORATORY - EUA)

Ponta Grossa, 30 de agosto de 2023.

Documento assinado eletronicamente por **Rafael Mazer Etto**, **Professor(a)**, em 31/08/2023, às 18:33, conforme Resolução UEPG CA 114/2018 e art. 1º, III, "b", da Lei 11.419/2006.

Documento assinado eletronicamente por **Wellington Claiton Leite**, **Usuário Externo**, em 04/09/2023, às 15:36, conforme Resolução UEPG CA 114/2018 e art. 1º, III, "b", da Lei 11.419/2006.

Documento assinado eletronicamente por **Jose Carlos Ferreira da Rocha**, **Coordenador(a) do Programa de Pós-Graduação em Computação Aplicada - Mestrado**, em 04/09/2023, às 16:35, conforme Resolução UEPG CA 114/2018 e art. 1º, III, "b", da Lei 11.419/2006.

A autenticidade do documento pode ser conferida no site https://sei.uepg.br/autenticidade informando o código verificador **1590195** e o código CRC **8B02112C**.

Dedico este trabalho aos meus pais, Marcos e Luiza.

AGRADECIMENTOS

Agradeço a Deus pela vida, oportunidade e perseverança que me concedeu para superar todos os obstáculos.

Aos meus pais, Marcos e Luiza, por todos os ensinamentos que me ajudaram a chegar até aqui e por estarem sempre ao meu lado, me motivando e apoiando nesta jornada.

À minha irmã Sayuri, pela nossa amizade e todo o carinho.

À minha esposa Paloma, por todo o apoio nos momentos adversos. Obrigado por todo o carinho, amor, companheirismo e compreensão. Você é o Estado da Arte.

Ao Prof. Dr. Rafael Mazer Etto, pela orientação, apoio, confiança e incentivo durante todo o processo de desenvolvimento deste trabalho.

À Profª. Drª. Carolina Weigert Galvão, por ser sempre solícita e auxiliar neste trabalho

Aos Me. Douglas Tomachewski e Me. Renann Rodrigues da Silva por todas as contribuições e apoio.

À Universidade Estadual de Ponta Grossa e ao programa de Pós-Graduação em Computação Aplicada por me receberem como discente e oferecerem todo o suporte necessário.

Aos professores do programa de Pós-Graduação em Computação Aplicada por dividirem o seu conhecimento.

Ao Laboratório de Biologia Molecular Microbiana pelo acolhimento.

Agradeço a todos que de alguma maneira contribuíram para o meu aprendizado nesta caminhada.

RESUMO

A identificação bacteriana é um tema de grande relevância no campo da agricultura para a compreensão da microbiologia do solo, sobretudo, a rizosfera. Entre as técnicas de identificação de microrganismos, a Espectrometria de Massa do tipo MALDI-TOF, do inglês *Matrix-Assisted Laser Desorption/Ionization Time-of-Flight*, tem sido extensivamente adotada como alternativa mais econômica e eficaz do que os métodos tradicionais, baseados em características fenotípicas. Este método permite a identificação de um microrganismo, pois cada microrganismo terá um perfil de espectro de massa único. No espectro de massa gerado, biomarcadores podem ser assinalados e utilizados como critério de classificação das amostras. Proteínas ribossomais são exemplos de biomarcadores confiáveis para a identificação bacteriana, pois desempenham funções vitais da célula e são altamente conservadas em sua sequência de aminoácidos. O classificador bacteriano Ribopeaks usa dados de massas moleculares dessas proteínas para a identificação do organismo. No entanto, espectros de bactérias inteiras podem conter picos correspondentes a proteínas não-ribossomais, peptídeos, metabólitos e lipídios em sua assinatura, o que dificulta a correta classificação. Diante disso, este trabalho utilizou da abordagem de agrupamento, por meio do algoritmo DBSCAN, do inglês *Density-Based Spatial Clustering of Applications with Noise*, para encontrar grupos de proteínas ribossomais sem distinção dos tipos, de modo a obter um filtro que identifique a partir da massa de uma macromolécula sua compatibilidade com uma proteína ribossomal. Para desenvolver os modelos, a base de dados Puchuy contendo as massas estimadas das proteínas ribossomais foi empregada. Essa base passou por uma fase de pré-processamento antes de ser submetida ao processo de aprendizado de máquina. Um automatizador para múltiplas classificações bacterianas no Ribopeaks foi construído para viabilizar a validação dos modelos gerados, enviando os organismos da base de dados reais SpectraBank ao classificador antes e após a filtragem dos picos. O filtro conseguiu aumentar discretamente a precisão do classificador bacteriano, ao mesmo tempo em que reduziu em média 40,1% do número de picos presentes na amostra bacteriana. Além disso, houve uma diminuição de 35,66% no tempo necessário para classificar esses mesmos organismos.

Palavras-chaves: Aprendizado de Máquina; Bioinformática; Biotecnologia; Espectrometria de Massa.

ABSTRACT

Bacterial identification is a topic of great standard in the field of agriculture for the understanding of soil microbiology, especially the rhizosphere. Among all techniques for identifying microorganisms, Mass Spectrometry MALDI-TOF type has been extensively adopted as a more economical and effective alternative than traditional methods, due to its phenotypic characteristics. This method facilitates microorganism identification, as each microorganism possesses a distinct mass spectrum profile. Within the produced mass spectrum, specific biomarkers can be assigned and utilized as criteria for sample classification. Ribosomal proteins are examples of biomarkers poised for bacterial identification, given their roles in cellular maintenance and their remarkable conservation in amino acid sequences. The Ribopeaks bacterial classifier uses molecular mass data from ribosomal proteins for organism identification. However, mass spectra data obtained from whole bacterial might include peaks associated with non-ribosomal proteins, peptides, metabolites, and lipids within their distinctive patterns, creating challenges for accurate classification. In this study, a clustering approach was employed, utilizing the DBSCAN algorithm, to cluster ribosomal proteins regardless of their specific types. This approach aimed to create a filter capable of determining the compatibility of a given macromolecule mass with a ribosomal protein. For construction of the models, the Puchuy base of presumed masses of ribosomal proteins was used, which went through a pre-processing step before being submitted to machine learning. A controller for multiple bacterial classifications in Ribopeaks was built to enable the validation of the generated models, sending the organisms from the real SpectraBank database to the classifier before and after filtering the peaks. In the best case, the filter was able to subtly increase the assertiveness of the bacterial classifier, with an average reduction of 40.1% in the peak volume of the bacterial sample and a reduction of 35.66% in the processing time for classification of the same organisms.

Palavras-chaves: Machine Learning; Bioinformatics; Biotechnology; Mass Spectrometry.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

API	*Application Programming Interface*
BPCV	Bactérias Promotoras do Crescimento Vegetal
DBSCAN	*Density-based spatial clustering of applications with noise*
EM	Espectrometria de Massa
JSON	*JavaScript Object Notation*
MALDI-TOF	*Matrix-assisted laser desorption/ionization time-of-flig*ht
NCBI	*National Center for Biotechnology*
SVM	*Support Vector Machine*
UV	Ultravioleta

SUMÁRIO

1 INTRODUÇÃO

O suprimento da demanda agrícola para alimentar a população humana estimada em 9,15 bilhões para 2050, simultaneamente ao aumento da produtividade para diminuir a pressão sobre o meio ambiente com a desaceleração da expansão de áreas de cultivo, é um dos desafios atuais da agricultura moderna (ALEXANDRATOS e BRUINSMA, 2012). Nesse contexto, as Bactérias Promotoras do Crescimento Vegetal (BPCV) presentes entre os microrganismos da rizosfera contribuem de forma benéfica para o crescimento das plantas, atuando como elicitores de tolerância à estresses bióticos, como agentes patogênicos (MAJEED; MUHAMMAD; AHMAD, 2018), e abióticos, como a salinidade do solo, estresse hídrico, entre outros fatores ambientais (KUMAR *et al*., 2020).

Estima-se que o número de espécies bacterianas por grama de solo seja de até 8,3 milhões (GANS; WOLINSKY; DUNBAR, 2005). No entanto, nem toda microbiota vegetal é benéfica, com microorganismos oportunistas que podem afetar a saúde das plantas e de humanos imunossuprimidos (MENDES; GARBEVA; RAAIJMAKERS, 2013). Dessa forma, a identificação das BPCV é um passo importante para utilizá-las com o intuito de aumentar a produtividade agrícola de maneira sustentável (SINGH *et al*., 2018).

O método de identificação de microorganismos tradicionalmente utilizado em laboratório é baseado em testes fenotípicos e geralmente demanda longos períodos de incubação. Como forma alternativa, a Espectrometria de Massa (EM) do tipo MALDI-TOF (ionização por dessorção a laser assistida por matriz - tempo de voo, do inglês *Matrix-assisted laser desorption/ionization time-of-flight*) demonstra maior eficiência quando comparada ao método convencional, com resultados mais confiáveis, melhor tempo de resposta e com média de custo oito vezes menor por amostra (LEGARRAGA *et al*., 2013).

A EM permite a identificação de uma amostra ao determinar sua massa em relação à sua carga, assim como a dos fragmentos gerados a partir dela (GARCÍA *et al*., 2012). As massas moleculares representados no espectro gerado podem ser analisados com o auxílio de aprendizado de máquina para a classificação da amostra, conforme os trabalhos de Tomachewski *et al*. (2018) e Silva (2021), que utilizam dados de proteínas ribossomais como biomarcadoras para a classificação de bactérias, devido à sua alta conservação (TERAMOTO *et al*., 2007). Entretanto, a discriminação dessas proteínas não é trivial, em razão das

sobreposições existentes em suas respectivas distribuições estatísticas observadas por Nascimento (2019).

Nesse sentido, este estudo investigou a abordagem de agrupamento no contexto da aprendizagem de máquina, utilizando o algoritmo DBSCAN para identificar grupos de proteínas ribossomais independentemente de seus tipos. O objetivo foi criar um filtro capaz de determinar, a partir da massa de uma macromolécula obtida por espectrometria de massa, a sua compatibilidade com uma proteína ribossomal.

Este trabalho está dividido nos seguintes capítulos: capítulo 2, objetivos, onde são definidos os objetivos geral e específicos para o trabalho. Capítulo 3, revisão de literatura, onde são abordados temas pertinentes para o desenvolvimento do trabalho. Capítulo 4, metodologia, onde são apresentadas as etapas de pré-processamento, geração dos modelos de agrupamento, desenvolvimento e validação do filtro. O capítulo 5 apresenta os resultados e discussões obtidas. O capítulo 6 apresenta as conclusões. Por fim, o capítulo 7 apresenta a publicação resultante da pesquisa.

2 OBJETIVOS

2.1 OBJETIVO GERAL

Desenvolver um filtro baseado em aprendizagem de máquina que identifique valores de massa/carga de proteínas ribossomais obtidos usando espectro de massa do tipo MALDI-TOF.

2.2 OBJETIVOS ESPECÍFICOS

- Construir um modelo de agrupamento baseado em massas moleculares estimadas de dados genômicos de proteínas ribossomais;
- Construir um automatizador para múltiplas classificações bacterianas no Ribopeaks;
- Avaliar o desempenho da associação do Ribopeaks com o filtro desenvolvido a partir do modelo gerado, utilizando dados completos reais de espectro de massa do tipo MALDI-TOF.

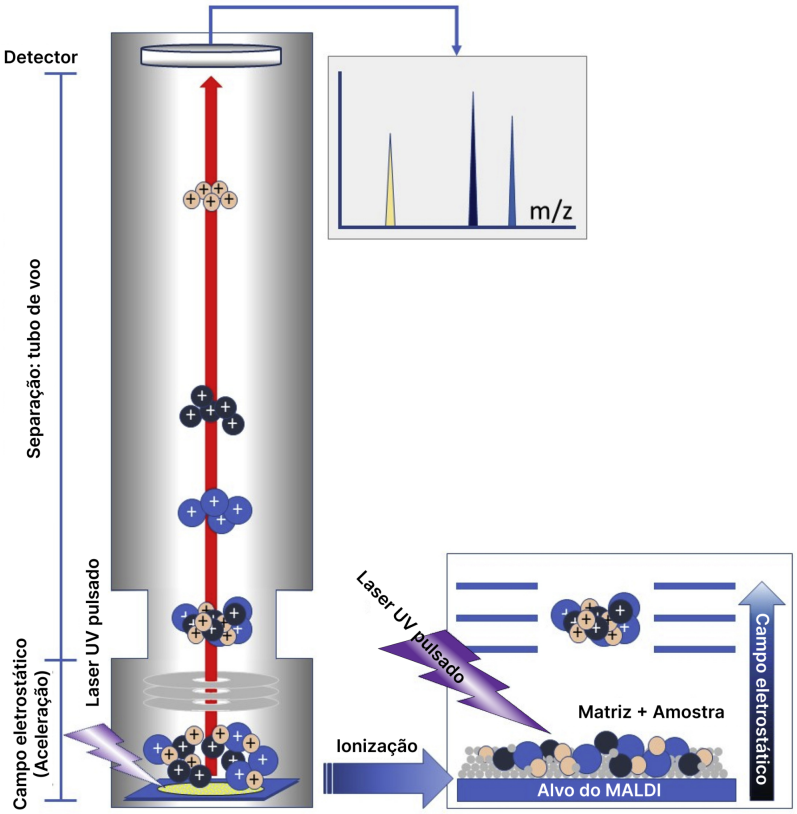
3 REVISÃO DE LITERATURA

3.1 ESPECTROMETRIA DE MASSA DO TIPO MALDI-TOF

A Espectrometria de Massa é uma técnica utilizada para gerar imagens de íons de amostras em valores da relação massa/carga (m/z), possibilitando mapear moléculas específicas para coordenadas bidimensionais (massa e intensidade relativa) da amostra original (CAPRIOLI; FARMER; GILE, 1997).

Na EM do tipo MALDI-TOF (Figura 1), a amostra é combinada com a matriz, que auxilia no processo de dessorção e ionização, e são depositadas em um suporte de amostra condutivo. Essa mistura passa por um processo de ionização quando exposta a um feixe de laser ultravioleta (UV). Isso resulta na extração e aceleração dos íons por meio de um campo elétrico, e esses íons são então transportados através de um tubo mantido a vácuo até o analisador de massas de tempo de voo. (BRONZEL JÚNIOR, 2015; JURINKE; OETH; VAN DEN BOOM, 2004).

Figura 1 - Representação do funcionamento do MALDI-TOF

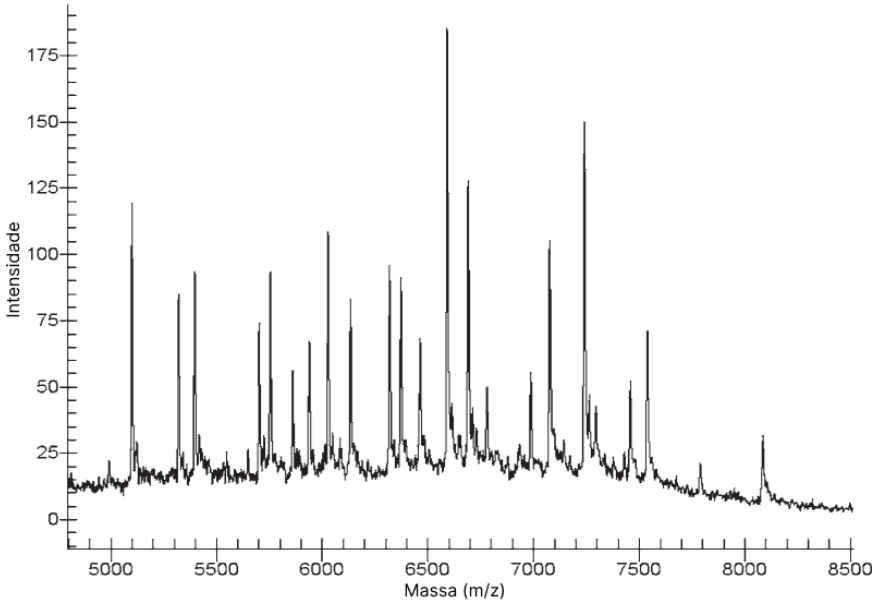


Fonte: Adaptado de Hou; Chiang-Ni e Teng (2019).

As moléculas ionizadas apresentam velocidades inversamente proporcionais à sua relação m/z. Dessa forma, moléculas de menor m/z alcançarão o detector mais rápido,

possuindo um tempo de voo mais curto quando comparado às maiores. O espectro resultante, ilustrado pela Figura 2, é composto pela intensidade do pico no eixo y e pelo tempo de voo no eixo x, que pode ser convertido para razão m/z usando valores conhecidos de m/z.

Figura 2 - Espectro de Massa



Fonte: Adaptado de Jurinke; Oeth e Van Den Boom (2004).

De acordo com Cuénod *et al.,* a EM do tipo MALDI-TOF revolucionou o diagnóstico microbiano e tornou-se o método de escolha para identificação de espécies bacterianas em diagnósticos clínicos, devido ao baixo custo, alta precisão e obtenção rápida do resultado.

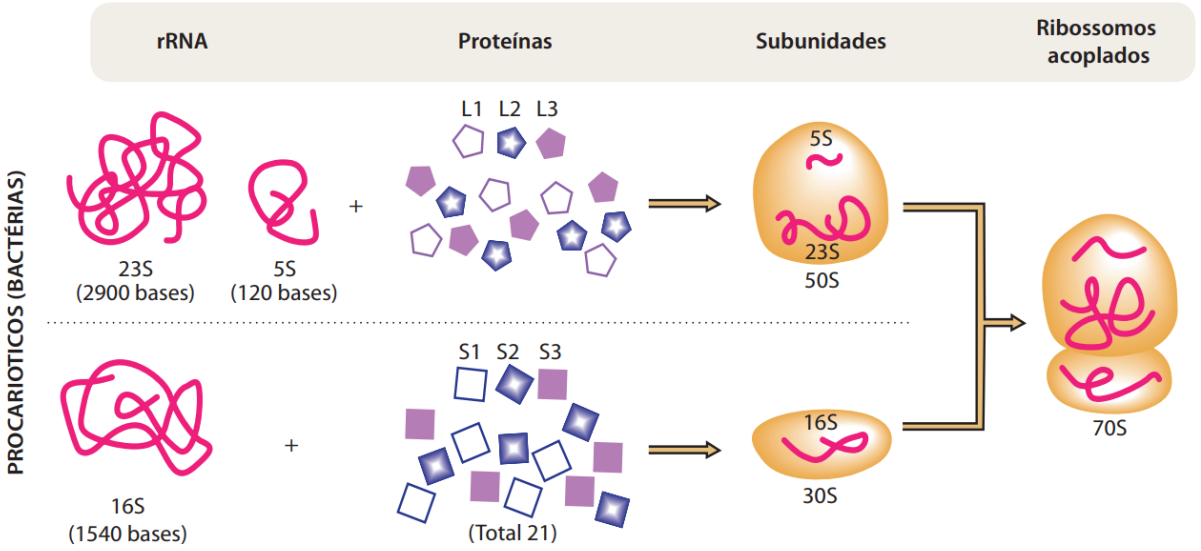
Um dos métodos de identificação bacteriana através de dados de EM do tipo MALDI-TOF é a comparação dos perfis espectrais com um banco de dados de referência previamente construído, como o banco de dados privado MALDI Biotyper 2.0 (Bruker Daltonics), considerado o maior banco, contendo mais de 1.800 espécies bacterianas, e o SpectraBank, uma base gratuita com mais de 200 cepas bacterianas e mais de 70 espécies bacterianas (BÖHME *et al.*, 2012).

No campo agrícola, a análise do perfil proteico obtido pelo MALDI-TOF tem sido utilizada na identificação de bactérias endofíticas, como as presentes nas raízes do alho (JUNIOR *et al.*, 2020) e na banana (MUTHURI *et al.*, 2012). Outra estratégia para a identificação bacteriana através do espectro de massa é a utilização de biomarcadores representados pelos picos de massa moleculares, como as proteínas ribossomais.

3.1.1 Proteínas Ribossomais

As proteínas ribossomais, junto com o rRNA, constituem o ribossomo, estrutura celular responsável pela síntese de proteínas. Em células procarióticas, o ribossomo é composto de duas subunidades de diferentes tamanhos, as subunidades maior (50S) e menor (30S) (RIBEIRO, 2009).

Figura 3 - Ribossomo procariótico



Fonte: Ribeiro (2009).

Essas proteínas desempenham funções vitais da célula e são altamente conservadas em sua sequência de aminoácidos, podendo ser consideradas biomarcadores confiáveis para a identificação bacteriana (TERAMOTO *et al.*, 2007), a exemplo de Ziegler *et al*. (2015), que utilizaram 13 proteínas ribossomais para caracterização de estirpes de BPCV. A abundância e onipresença dessas proteínas e a sua capacidade de ionizar (DE CAROLIS *et al.*, 2014), favorece a sua visualização no MALDI-TOF.

3.2 APRENDIZADO DE MÁQUINA NÃO-SUPERVISIONADO

De acordo com Lorena; Gama e Faceli (2000), o método de aprendizagem não-supervisionada faz o uso de algoritmos inteligentes para encontrar padrões e relações em uma base com dados sem rotulação prévia, com o objetivo de estabelecer a existência de classes ou agrupamentos (*clusters*). Esta técnica é utilizada quando, para cada instância,

apenas atributos de entrada estão disponíveis. Dessa forma, esses padrões ou tendências auxiliam no entendimento dos dados (DE SOUTO *et al.,* 2003).

O agrupamento é, portanto, uma técnica de aprendizado de máquina não-supervisionado que busca encontrar uma estrutura em uma coleção de dados não rotulados, com base em similaridades ou dissimilaridades entre os padrões (MADHULATHA, 2012). Algoritmos de agrupamento são popularmente conhecidos por serem eficientes na resolução de problemas de mineração de dados, como o DBSCAN (Clusterização Espacial Baseada em Densidade de Aplicações com Ruído, do inglês *Density-based spatial clustering of applications with noise*) (AHMED; SERAJ; ISLAM, 2020; SCHUBERT *et al.,* 2017).

3.2.1 DBSCAN

O DBSCAN é um algoritmo projetado para descobrir *clusters* de formatos arbitrários e tamanhos distintos para bases que possuam dados de um espaço métrico (ESTER *et al.*, 1996). A ideia-chave do algoritmo determina que, para cada ponto dentro de um *cluster*, a densidade na vizinhança deve ultrapassar um limite específico, ou seja, deve haver um número mínimo de pontos dentro de um determinado raio.

A forma de uma vizinhança é determinada pela escolha de uma função de distância (como a distância euclidiana) para um par de pontos p e q pertencentes ao domínio finito D, denotada por $dist(p, q)$. A vizinhança de raio Eps de um ponto p, denotada por $NEps(p)$, é definida como:

$$NEps(p) = \{q \in D \mid dist(p, q) \leq Eps\}$$

Existem dois tipos de pontos em um *cluster*: pontos centrais (dentro do *cluster*) e pontos de borda (na borda do *cluster*). Dessa forma, para cada ponto central p em um *cluster* C, deve haver um ponto q em C de modo que p esteja dentro da vizinhança de raio Eps de q e que $NEps(p)$ satisfaça a quantidade mínima de pontos definida. Por outro lado, caso a vizinhança de raio Eps de um ponto não atinja a quantidade mínima mas contenha algum ponto central, este será considerado um ponto de borda.

Como exemplo no contexto da microbiologia do solo, o DBSCAN foi utilizado para extrair e analisar características gerais da morfologia de colônias a fim de discriminar fenótipos distintos de *Bacillus subtilis* (MAYER; HOLTRUP; GRAUMANN, 2022).

3.3 TRABALHOS CORRELATOS

Hotta *et al.* (2010) propõem o método rápido de identificação bacteriana por EM do tipo MALDI-TOF através da criação de um banco de dados baseado em proteínas ribossomais do operon S10-spc-alpha como biomarcadoras, calculando as massas de acordo com suas sequências de aminoácidos. Posteriormente, esta técnica foi utilizada no trabalho apresentado por Tamura, Hotta e Sato (2013), tendo sua eficácia evidenciada na identificação da estirpe *Pseudomonas syringae*, agente patogênico causador de doenças em plantas. Ainda na esfera agrícola, as proteínas ribossomais também se mostraram suficientes na diferenciação de estirpes comuns de bactérias promotoras do crescimento vegetal (ZIEGLER *et al.,* 2015).

Tomachewski (2017) desenvolveu a base de dados PUKYU, incorporando massas moleculares estimadas para as 60 proteínas ribossomais frequentemente encontradas nas subunidades 50S e 30S do ribossomo. As sequências dos aminoácidos obtidas no NCBI (Centro Nacional de Informação Biotecnológica, do inglês *National Center for Biotechnology Information*) foram convertidas em pesos moleculares estimados através de uma calculadora desenvolvida. Ao todo, a base de dados abrange 1.949 gêneros e 6.936 espécies únicas, totalizando 28.505 registros, que incluem também informações ausentes.

A base PUKYU foi empregada no desenvolvimento do Ribopeaks (Tomachewski *et al.*, 2018), um *software* de classificação de bactérias que recebe uma entrada de dados de m/z geradas por MALDI-TOF e e utiliza um modelo gerado pelo classificador probabilístico Naïve Bayes para determinar a classificação taxonômica do conjunto inserido com base nas probabilidades calculadas. Ao ser testado com dados extraídos de Ziegler *et al.* (2015), o Ribopeaks obteve as taxas de acerto de 90,51% e 87,93% para classificações taxonômicas a nível de gênero e espécie, respectivamente.

Silva (2021), por sua vez, extraiu do repositório NCBI apenas dados de bactérias com genoma completamente sequenciado para obtenção dos dados genômicos de proteínas ribossomais para compor a base de dados Puchuy. O cálculo das massas moleculares ocorreu de forma semelhante ao de Tomachewski *et al.* (2018), com uma versão adaptada em linguagem Python da calculadora de peptídeos. Após o pré-processamento com a padronização dos rótulos, foi realizado o treinamento do *ensemble* baseado em agrupamento.

Os algoritmos do tipo máquina de vetores de suporte (SVM, do inglês *Support Vector Machine*), *Decision Tree* e *Random Forest* obtiveram os melhores resultados. Entretanto,

todos eles possuem limitações. O SVM, por ser um método computacionalmente custoso, não pôde ser executado em níveis taxonômicos de Família, Gênero e Espécie. Por outro lado, os algoritmos *Decision Tree* e *Random Forest* são prejudicados em casos reais onde haja dados faltantes. Todavia, os resultados são positivos, alcançando um ganho de 9% no desempenho no melhor caso com a utilização de agrupamento.

A classificação dos biomarcadores identificados no Espectro de Massa é comumente realizada através da comparação com valores estimados de produtos gênicos correspondentes à massa medida (LAUBER; RUNNING; REILLY, 2009; ZAUTNER *et al*., 2015).

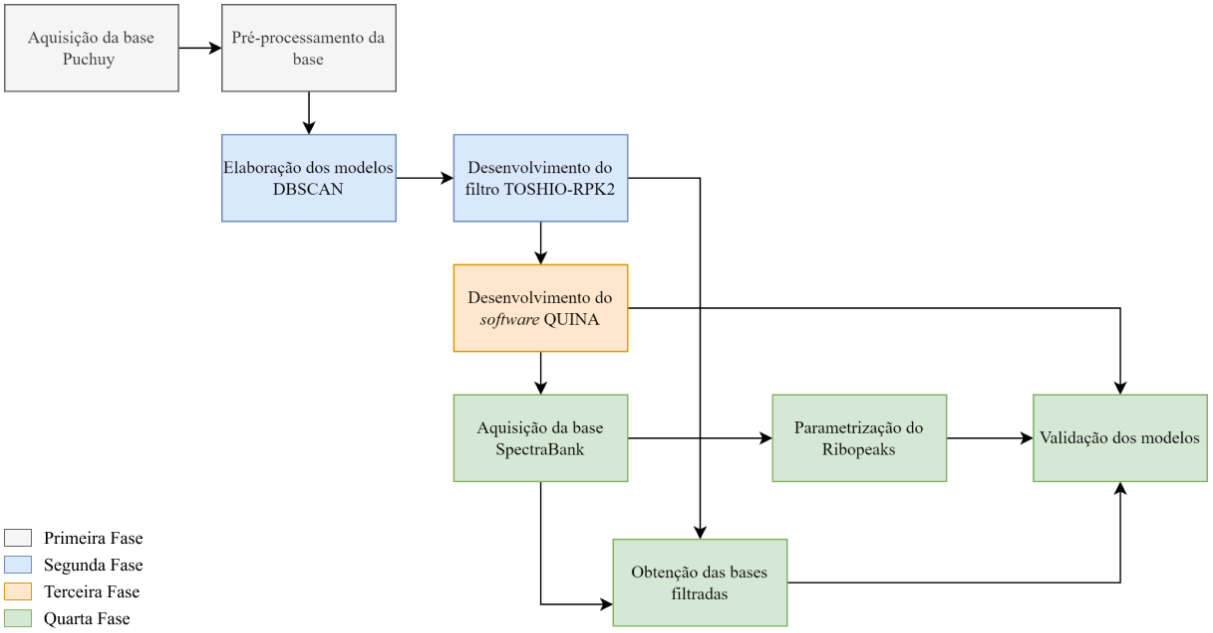
Nascimento (2019) associa o conjunto de m/z de cada uma das proteínas ribossomais presentes na base de dados PUKYU com distribuições estatísticas para criar um modelo classificatório e representativo a fim de discriminá-las a partir do espectro de massa MALDI-TOF. A partir dos histogramas gerados, o autor observou que em algumas proteínas, como a L12 e a L17, apresentavam mais de um pico em sua distribuição, indicando a presença de características singulares em suas distribuições.

Com isso, foi aplicado o modelo de mistura com o algoritmo *Partitioning Around Medoids* para a identificação dos subgrupos, já que a complexidade desses conjuntos não poderia ser compreendida através de simples distribuições estatísticas. Para verificar o ajuste dos dados de cada proteína às distribuições, foi utilizado o teste estatístico Cramér-Von Mises, conhecido como teste de significância ou teste de hipóteses. Os resultados obtidos no uso do Algoritmo Genético para selecionar o melhor modelo classificatório para as proteínas ribossomais demonstraram uma alta taxa de sobreposição das suas distribuições estatísticas.

4 METODOLOGIA

Quatro fases foram estabelecidas para a realização do projeto, demonstradas no fluxograma da figura 4. Na primeira fase, a base de dados de massas estimadas de proteínas ribossomais foi obtida e pré-processada para uso no treinamento do algoritmo de aprendizado de máquina. Na segunda fase, foram gerados quinze variantes do modelo de agrupamento DBSCAN e o filtro para a identificação das proteínas foi construído. Na terceira fase, um automatizador para múltiplas classificações bacterianas no Ribopeaks foi construído para viabilizar a validação dos modelos gerados. Por fim, na quarta fase uma base contendo dados reais de organismos inteiros foi obtida para avaliar o desempenho da associação do Ribopeaks com o filtro desenvolvido a partir dos modelos gerados.

Figura 4 - Fluxograma da metodologia



Fonte: o autor.

4.1 AQUISIÇÃO E PRÉ-PROCESSAMENTO DA BASE DE DADOS DE PROTEÍNAS RIBOSSOMAIS

A base de dados Puchuy (SILVA, 2021), selecionada para o treinamento dos algoritmos e disponibilizada pelo autor, é composta por dados presumidos de massa/carga (m/z) de proteínas ribossomais de bactérias, contendo 14.689 registros distribuídos entre

1.163 gêneros e 3.253 espécies. Cada registro da base possui 59 atributos que correspondem às massas moleculares e um atributo meta que descreve a taxonomia da bactéria. As proteínas ribossomais que compõem este conjunto de atributos são: L1, L2, L3, L4, L5, L6, L7A, L10, L11, L7/L12, L13, L14, L15, L18, L22, L23, L24, L29, L30, S2, S3, S4, S5, S7, S8, S9, S10, S11, S12, S13, S14, S15, S17, S19, L7ae, L9, L16, L17, L19, L20, L21, L25, L27, L28, L31, L32, L33, L34, L35, L36, S1, S6, S16, S18, S20, S21, S22, THX e YCF65.

Para a construção da base, Silva (2021) calculou as massas moleculares das proteínas a partir de dados de genomas completos extraídos do repositório NCBI utilizando uma versão adaptada da calculadora de peptídeos que considera as modificações pós-traducionais, proposta por Tomachewski *et al.* (2018). Ao todo, a base é constituída de 788.403 dados de m/z.

4.1.1 Pré-processamento da base de dados Puchuy

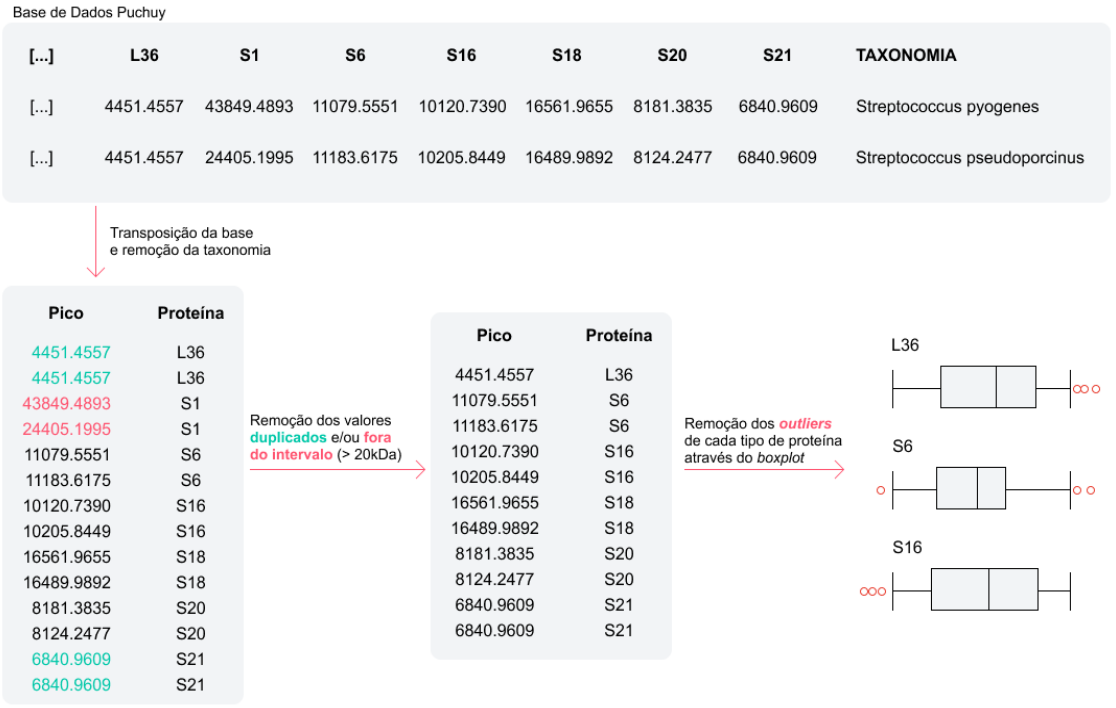
Para o desenvolvimento do filtro de identificação de proteínas ribossomais, a taxonomia da amostra não foi considerada na análise. Dessa forma, a base de dados elaborada por Silva (2021) foi adaptada para a execução deste trabalho com o auxílio de *scripts* implementados em linguagem Python..

A primeira etapa do pré-processamento da base de dados Puchuy consistiu na redução de dimensionalidade. Nesta etapa, a coluna correspondente ao atributo de taxonomia foi removida e os dados foram transpostos de modo a formar uma base para cada tipo de proteína contendo dois atributos: *pico*, representando o valor da massa molecular, e *proteína*, informando a proteína ribossomal correspondente.

Simultaneamente, foram desconsideradas as instâncias cujos picos de massa não pertenciam ao intervalo de 2.000 a 20.000 Da, amplitude na qual as cepas bacterianas codificam um número variável de marcadores ribossomais visíveis nos espectros de massa gerados pelo MALDI-TOF (ESTIGARRIBIA, 2017; CUÉNOD, 2021). Ainda, foram removidos os valores duplicados de modo a dirimir o viés da base.

Posteriormente, foi gerado o *boxplot* para cada proteína ribossomal, que permitiu a visualização de suas distribuições e identificação e remoção dos valores atípicos. Por fim, os dados das proteínas ribossomais foram unificados em uma única base para treinamento do modelo. A figura 5 ilustra a fase de pré-processamento.

Figura 5 - Pré-processamento da base



Fonte: o autor.

A base resultante é composta de 142.827 registros de kDa presumidos. Os quantitativos das proteínas constam na Tabela 1.

Tabela 1 - Distribuição das proteínas

L1	L2	L3	L4	L5	L6	L7A	L7AE	L7L12	L9
3	2	6	4	510	4136	41	541	3915	3626
L10	**L11**	**L13**	**L14**	**L15**	**L16**	**L17**	**L18**	**L19**	**L20**
4024	1989	3771	2704	3874	3243	3529	3672	3503	3369
L21	**L22**	**L23**	**L24**	**L25**	**L27**	**L28**	**L29**	**L30**	**L31**
2972	3289	3315	3278	509	3324	3371	3050	2661	4120
L32	**L33**	**L34**	**L35**	**L36**	**S1**	**S2**	**S3**	**S4**	**S5**
3181	2768	1974	3039	1828	5	1	6	6	2204
S6	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**	**S15**
3449	3239	3572	3791	2044	2741	1802	3451	3613	3485
S16	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**THX**	**YCF65**	
3750	3302	4743	2744	3460	2007	11	201	59	

Fonte: o autor.

4.2 ELABORAÇÃO DOS MODELOS DE AGRUPAMENTO E DESENVOLVIMENTO DO FILTRO

Após a etapa de pré-processamento, a base de dados de massas presumidas de proteínas ribossomais foi utilizada para treinamento do algoritmo de classificação não supervisionada DBSCAN, visando a criação de agrupamentos com os dados dos picos de m/z. Para tal, a implementação do modelo foi realizada em linguagem Python com auxílio da biblioteca Scikit-learn (PEDREGOSA et al., 2011).

A biblioteca disponibiliza dois parâmetros para ajuste da densidade dos *clusters*: *"min_samples"*, que representa a quantidade mínima de amostras em uma vizinhança para a definição de um *cluster* e *"eps"*, que representa a distância máxima entre duas amostras para serem consideradas vizinhas pertencentes a um mesmo *cluster*.

Foram gerados 15 modelos distintos utilizando o algoritmo DBCAN com o parâmetro *"min_samples"* definido no valor "1" para preservação de todos os picos da base, com o parâmetro *"eps"* incrementado em 0,1 dentro do intervalo de 0,1 a 1,5 Da para ajuste da abrangência dos *clusters*. Por ser um problema de abordagem unidimensional, a métrica utilizada para o cálculo da diferença absoluta entre dois picos x e y foi a distância Euclidiana, dada pela seguinte equação:

$$d(x, y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$$

Nesta etapa, o filtro nomeado TOSHIO-RPK2 (*Tool of Selection Highly Operationalized for RiboPeaKs2*) foi desenvolvido na linguagem Python para identificação de massas compatíveis com proteínas ribossomais a partir da entrada de um conjunto de valores da relação massa/carga (m/z) de uma amostra bacteriana, obtidos por Espectrometria de Massa do tipo MALDI-TOF ou cálculo de massa molecular de peptídeos por meio de dados genômicos.

Este filtro carrega um modelo DBSCAN previamente treinado a partir de um arquivo usando a biblioteca *pickle*. Em seguida, solicita ao usuário que insira os picos do organismo separados por ponto e vírgula. Os valores inseridos pelo usuário são transformados em um vetor do tipo *array numpy* e enviados para a função "*dbscan_predict*", usada para prever a qual cluster cada pico pertence com base no modelo carregado. A função calcula a distância

entre cada pico e os componentes centrais do modelo DBSCAN e atribui o rótulo do cluster mais próximo se a distância for menor que um valor de *eps*. Finalmente, o código imprime os picos selecionados com seus respectivos rótulos de cluster ou informa que nenhum pico foi selecionado, dependendo do resultado do processo.

4.3 DESENVOLVIMENTO DO SOFTWARE QUINA

A classificação de bactérias por meio de dados da relação m/z de proteínas ribossomais através do Ribopeaks (Tomachewski *et al.*, 2018) é realizada com a submissão dos valores a serem analisados em um campo da interface *web*, que retorna os organismos correspondentes aos dados inseridos. Para viabilizar os testes de classificação de múltiplos microrganismos, o *software* nomeado QUINA (Automatizador para classificação bacteriana em lote no Ribopeaks) foi desenvolvido.

O *software,* desenvolvido em linguagem Python, seleciona cada organismo da base informada e envia ao classificador por meio de uma API (Interface de Programação de Aplicação, do inglês *Application Programming Interface*), formata o arquivo JSON (*JavaScript Object Notation*) retornado pelo servidor e organiza as informações de classificação e métricas adicionais da requisição (número de picos enviados, tempo de execução da análise, taxonomia da bactéria enviada, informações de êxito na classificação e lista de bactérias correspondentes com as respectivas proteínas ribossomais selecionadas para a predição), agrupando todas as classificações da base em um único arquivo de texto no armazenamento local.

4.4 VALIDAÇÃO DOS MODELOS DE AGRUPAMENTO DBSCAN

Os testes realizados com dados de proteínas ribossomais extraídos de Ziegler *et al.* (2015) por Tomachewski *et al.* (2018) no Ribopeaks demonstraram a alta precisão na classificação das 116 cepas bacterianas ambientais. Contudo, um espectro de uma bactéria inteira pode conter picos correspondentes a proteínas não-ribossomais, peptídeos, metabólitos e lipídios em sua assinatura, o que dificulta a correta classificação.

A base de dados SpectraBank (BÖHME *et al.*, 2012), selecionada para a avaliação do desempenho da associação do filtro desenvolvido com o Ribopeaks, disponibiliza de forma

gratuita espectros e listas de picos de massas moleculares para mais de 200 cepas bacterianas e mais de 70 espécies bacterianas. Os espectros de massa das amostras foram obtidos em uma faixa de m/z de 1.500 a 15.000 Da com a utilização do equipamento Voyager DE STR MALDI-TOF (Applied Biosystems).

Como meio de parametrização do software, o Ribopeaks permite o ajuste do pré-definido erro de tolerância de massa de 3 Da, calibrado através dos testes anteriormente citados. Dessa forma, para obter o melhor cenário de classificação a nível de Gênero e Espécie com dados do SpectraBank, a base foi submetida sucessivas vezes no classificador com o auxílio do QUINA, variando o erro de tolerância de massa com incremento de 1 Da dentro do intervalo de 1 a 10 Da.

Por fim, o SpectraBank foi submetido ao filtro TOSHIO-RPK2 junto aos modelos elaborados para seleção dos picos de cada amostra, resultando em quinze bases com registros filtrados. Todas elas tiveram seus organismos encaminhados para classificação no Ribopeaks através do QUINA

As análises realizadas baseiam-se nas métricas de acurácia para ambos os níveis taxonômicos, tempo de classificação total e quantidade de picos enviados.

5 RESULTADOS E DISCUSSÃO

A base de treinamento, obtida do pré-processamento da base Puchuy, é constituída de 142.827 registros de picos de proteínas ribossomais dos 788.403 originalmente disponibilizados. Ao ordenar os dados de forma crescente, foi observado que os tipos de proteínas se alternam 100.629 vezes. No entanto, apenas em duas ocorrências a diferença de massa entre proteínas vizinhas é igual ou maior que 10 Da. A média dessa distância em todas as alternâncias é de 0,116 ± 0,192 Da. Nesta ordenação, a proteína L5 possui o maior número de picos de valores distintos agrupados sequencialmente sem interrupção, com 333 registros.

A Tabela 2 apresenta os valores calculados para a elaboração do *boxplot* das cinco proteínas mais populosas da base. O elevado valor do intervalo interquartil (IIQ) constata a alta dispersão dos dados dentro do intervalo trabalhado. Essa dispersão colabora com o surgimento de sobreposições das faixas de abrangência das proteínas como observado pelo primeiro quartil (Q1) e terceiro quartil (Q2) das proteínas S18 e L15, dificultando a identificação individual. A relação completa encontra-se no apêndice A e os *boxplots* encontram-se no apêndice B.

Tabela 2 - Valores dos quartis das proteínas mais populosas

Proteína	Quantidade	Outliers	Q1	Q2	Q3	IIQ
S18	4743	0	10046.8088	16482.6518	17333.4756	7286.6668
L31	4171	51	7979.0258	8980.1735	9744.6422	1765.6163
L6	4144	8	18923.0385	19197.9136	19419.7541	496.7156
L15	4074	200	15230.9255	15626.61805	16199.9336	969.0081
L7L12	4033	118	12434.1781	12654.4335	13047.0478	612.8697

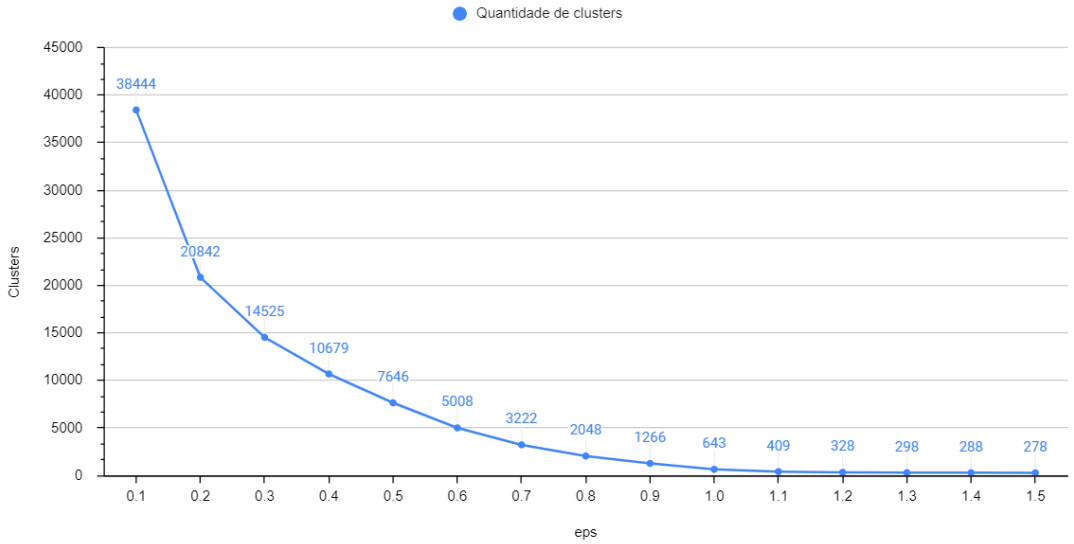
Fonte: o autor.

Ainda, foram observados 100 valores de m/z compartilhados por duas proteínas distintas, os quais estão discriminados no apêndice C. Tais resultados expuseram a sobreposição das proteínas no intervalo de análise, o que motivou a utilização da abordagem de aprendizado de máquina não-supervisionada.

O DBSCAN é um algoritmo de agrupamento capaz de se moldar aos dados unidimensionais utilizando a distância entre pontos como critério principal para definir regiões densas. Através dele, foram elaborados 15 modelos com variação do epsilon para

obter o melhor ajuste do tamanho dos clusters. O número de clusters para cada modelo variou de 38.444 a 278, ilustrado pela Figura 6.

Figura 6 - Quantidade de *clusters* com a variação do *eps*

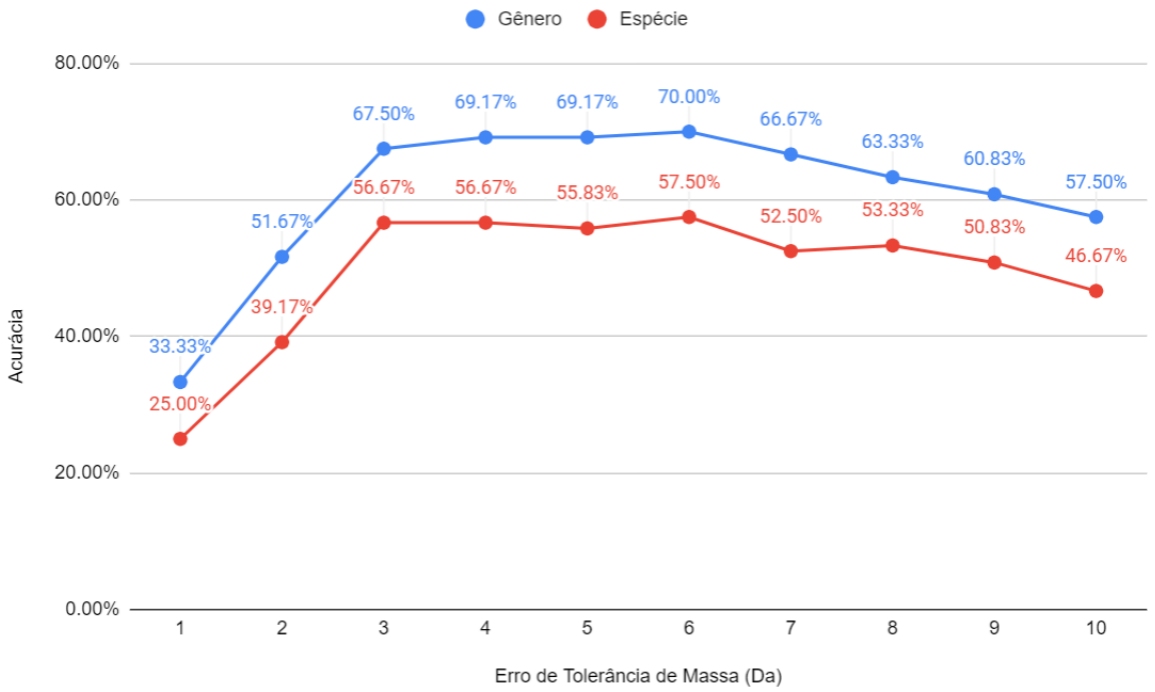


Fonte: o autor.

O cluster mais populoso no modelo gerado com o parâmetro eps=0,9 contém 29.444 picos e abrange dados de 24 proteínas. Além disso, foi observado que 676 clusters nesse modelo exibem a presença de dois ou mais tipos de proteínas, indicando a existência de uma mistura de grupos distintos de proteínas dentro de uma mesma vizinhança.

Através da parametrização do Ribopeaks, conforme descrito na seção 4.4 e representado pela Figura 7, evidenciou-se que a maior acurácia para ambos os níveis taxonômicos com a submissão de toda a base SpectraBank foi obtida com a definição de 6 Da para o erro de tolerância de massa, sendo 70,00% a nível de Gênero e 57,50% a nível de Espécie. O tempo total de execução para classificação de todos os organismos foi de 15.794,28 segundos.

Figura 7 - Ajuste do erro de tolerância de massa do Ribopeaks



Fonte: o autor.

Após a parametrização do Ribopeaks, os modelos gerados foram aplicados de forma individual para filtrar a base SpectraBank. As bases filtradas resultantes, ou seja, apenas com potenciais dados de proteínas ribossomais, foram submetidas ao Ribopeaks através do QUINA utilizando 3 e 6 Da de erro de tolerância de massa na classificação.

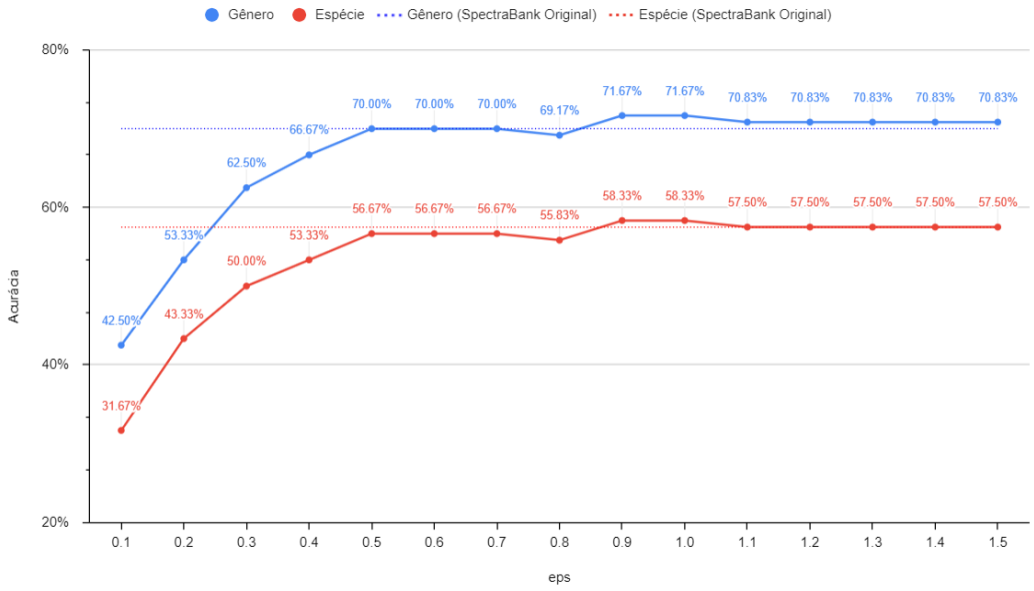
Quanto maior a abrangência dos clusters, mais picos são selecionados pelo filtro até que todos os registros da base sejam considerados. A Figura 8 demonstra a progressão da quantidade de picos selecionados da base original conforme cada modelo gerado, representado pelo seu atributo “eps”. A linha vermelha do gráfico representa os 2.758 picos totais contidos no SpectraBank. A relação detalhada da quantidade de picos selecionados por organismo encontra-se no apêndice D.

Figura 8 - Quantidade de picos selecionados pelo filtro

Fonte: o autor.

Os resultados de assertividade na classificação das bases filtradas foram avaliados a nível de Gênero e Espécie. O cenário mais otimizado foi alcançado com o modelo configurado para 0,9 Da e 1,0 Da de distância entre dois picos para serem considerados pertencentes a um mesmo cluster. Utilizando esse filtro, a precisão atingiu 71,67% a nível de Gênero e 58,33% a nível de Espécie. Notavelmente, esses valores superaram ligeiramente aqueles valores obtidos com o envio do SpectraBank original ao classificador bacteriano parametrizado com 6 Da de erro de tolerância de massa, conforme mostra a Figura 9.

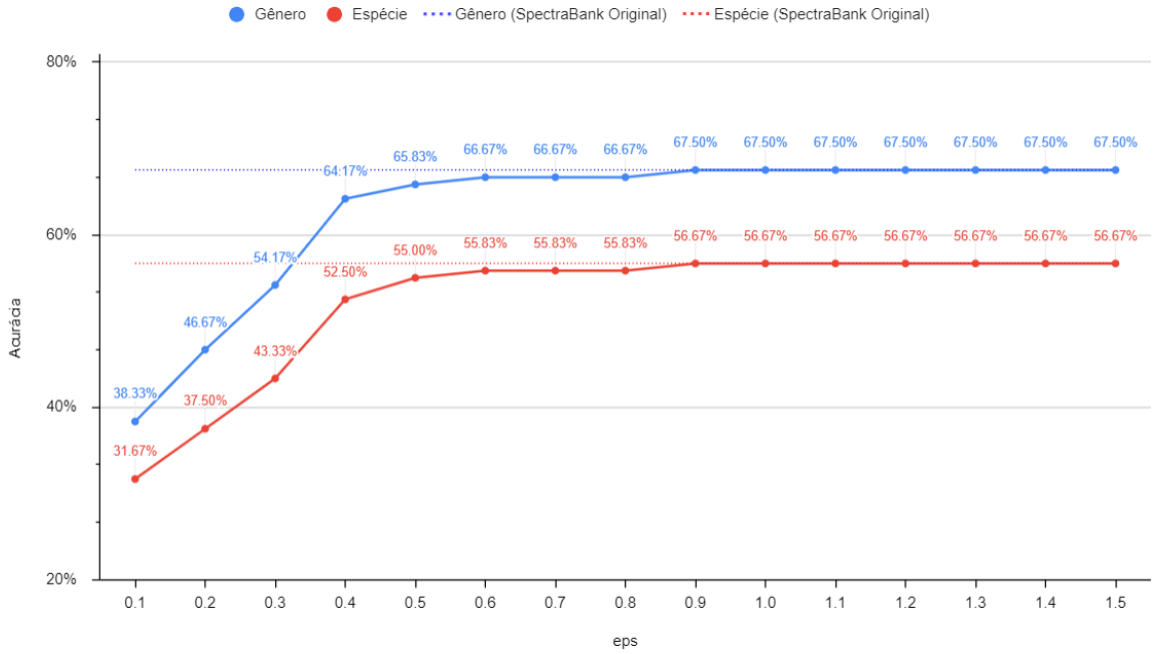
Figura 9 - Classificação das bases filtradas com 6 Da de erro de tolerância



Fonte: o autor.

Levando em consideração que o filtro possui como objetivo a seleção de proteínas ribossomais e que o Ribopeaks foi calibrado para trabalhar com esses dados, as bases filtradas foram submetidas ao classificador utilizando-se o valor padrão de erro de tolerância de massa de 3 Da. No entanto, a assertividade do Ribopeaks com o filtro não ultrapassou a obtida com o envio do SpectraBank completo sob as mesmas condições, de 67,5% a nível de Gênero e 56,67% a nível de Espécie, como mostrado na Figura 10.

Figura 10 - Classificação das bases filtradas com 3 Da de erro de tolerância



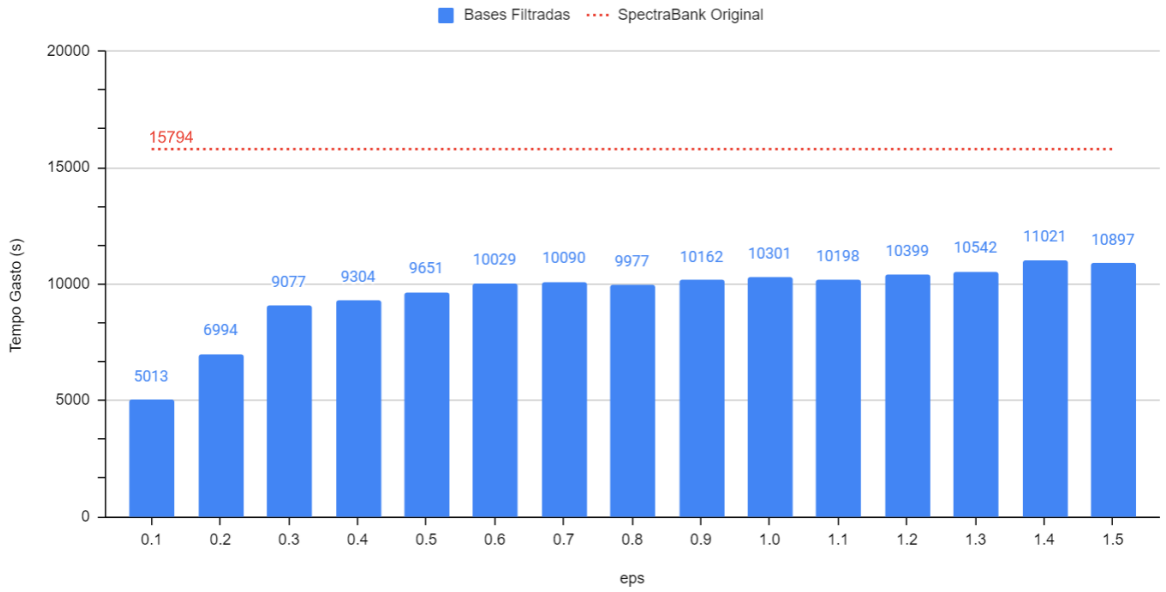
Fonte: o autor.

Um fator a ser considerado na avaliação do desempenho do filtro baseados nos modelos de agrupamento é a manutenção das classificações corretas dos organismos com o envio do SpectraBank original. Nos melhores casos obtido com o Ribopeaks parametrizado em 6 Da, todas as classificações corretas foram preservadas, acertando adicionalmente a bactéria *Carnobacterium divergens* ATCC 35677 a nível de Gênero e Espécie, juntamente com *Carnobacterium gallinarum* ATCC 49517 a nível de Gênero.

Com a utilização do filtro com o modelo modelo de eps = 0,9 foram desconsiderados 1.106 picos, o que representa 40,1% de redução do volume da base. O envio de menos picos por organismo reflete no tempo necessário para a classificação. A Figura 11 apresenta o

tempo gasto em segundos para a classificação de cada base no Ribopeaks parametrizado em 6 Da.

Figura 11 - Tempo gasto para classificação das bases filtradas



Fonte: o autor.

Quanto ao tempo necessário para classificar todos os organismos presentes no SpectraBank, a utilização do Ribopeaks com a parametrização de 6 Da levou 10.162 segundos para a base contendo organismos filtrados pelo modelo com eps=0,9. Comparativamente, foram requeridos 15.794 segundos para classificar a base original do SpectraBank. Isso representa uma notável redução de 35.66% no tempo de processamento.

Tabela 3 - Proteínas presentes em classificações corretas

L36	**L30**	**L34**	**L29**	**L32**	**L35**	**S21**	**L33**	**L28**
90,80%	72,41%	65,52%	60,92%	59,77%	42,53%	33,33%	31,03%	25,29%
L31	**S20**	**L27**	**S15**	**S22**	**S14**	**S16**	**L9**	
17,24%	17,24%	5,75%	4,60%	3,45%	2,30%	2,30%	1,15%	

Fonte: o autor.

Por fim, a Tabela 3 apresenta a relação das proteínas utilizadas pelo Ribopeaks em classificações corretas a nível de Gênero e Espécie. A proteína L36 foi utilizada em 90,8% das classificações corretas, o que pode indicar uma potencial proteína biomarcadora.

6 CONCLUSÃO

Neste trabalho avaliou-se o uso de técnica de aprendizado de máquina para a identificação de proteínas ribossomais em espectro de massa do tipo MALDI-TOF. Para treinamento do algoritmo, utilizou-se uma base de dados presumidos de massa/carga (m/z) de proteínas ribossomais de bactérias. A base de dados de validação, por sua vez, é constituída de dados reais de amostras bacterianas obtidos através de Espectrometria de Massa.

A abordagem de agrupamento utilizada através do algoritmo DBSCAN evidenciou as sobreposições de grupos distintos de proteínas dentro do intervalo de 2.000 a 20.000 Da, amplitude na qual as cepas bacterianas codificam um número variável de marcadores ribossomais visíveis nos espectros de massa gerados pelo MALDI-TOF. Essa distribuição dificulta a distinção entre as diferentes proteínas.

O modelo gerado com a definição de 0,9 Da como distância máxima entre duas amostras para serem consideradas vizinhas dentro de um mesmo *cluster* demonstrou resultados positivos ao filtrar os picos de espectros reais, reduzindo o ruído dos dados e otimizando o tempo de processamento para classificação pelo Ribopeaks.

Este trabalho resultou em um filtro de proteínas ribossomais que reduz em média 40% do volume de picos de uma amostra bacteriana, otimizando o processo de classificação pelo Ribopeaks. No entanto, não há garantia que alguns picos correspondentes a proteínas ribossomais não sejam desconsiderados durante a análise do filtro.

Os resultados obtidos com o uso do algoritmo DBSCAN para agrupamento das proteínas são promissores, elevando a assertividade do classificador bacteriano para os casos testados. Novas abordagens para a elaboração dos modelos podem ser avaliadas em trabalhos futuros, como a seleção de proteínas ribossomais biomarcadoras para composição da base de treinamento, consideração de deslocamento de picos ocasionados pela resistência à antibióticos e a unificação de modelos de proteínas individuais para uso no filtro.

7 PUBLICAÇÕES RESULTANTES DA PESQUISA

Esta pesquisa gerou um registro de programa de computador do *software* descrito no capítulo 4.3.

Patente: Programa de Computador

Número do registro: BR512022002873-5

Data de registro: 25/10/2022

Título: QUINA (Automatizador para classificação bacteriana em lote no Ribopeaks)

Instituição de registro: INPI - Instituto Nacional da Propriedade Industrial

Adicionalmente, um registro de programa de computador para o filtro descrito no capítulo 4.2.

Patente: Programa de Computador

Número do registro: BR512023002262-4

Data de registro: 08/08/2023

Título: TOSHIO-RPK2 - Tool Of Selection HIghly Operationalized for RiboPeaKs2

Instituição de registro: INPI - Instituto Nacional da Propriedade Industrial

.

REFERÊNCIAS

AHMED, M.; SERAJ, R.; ISLAM, S. M. S. The k-means algorithm: A comprehensive survey and performance evaluation. **Electronics**, v. 9, n. 8, p. 1295, 2020.

ALEXANDRATOS, N.; BRUINSMA, J. World agriculture towards 2030/2050: the 2012 revision. 2012.

BÖHME, K. et al. Spectra Bank: An open access tool for rapid microbial identification by MALDI‐TOF MS fingerprinting. **Electrophoresis**, v. 33, n. 14, p. 2138-2142, 2012.

BRONZEL JÚNIOR, J. L.. **Matrizes iônicas: detecção e quantificação de cianotoxinas por maldi-ms**. 2015. Dissertação (Mestrado em Química) - Universidade Estadual Paulista Júlio de Mesquita Filho, Araraquara, 2015.

CAPRIOLI, R. M.; FARMER, T. B.; GILE, J. Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS. **Analytical chemistry**, v. 69, n. 23, p. 4751-4760, 1997.

CUÉNOD, A. et al. Factors associated with MALDI-TOF mass spectral quality of species identification in clinical routine diagnostics. **Frontiers in cellular and infection microbiology**, v. 11, p. 104, 2021.

DE CAROLIS, E. *et al*. Application of MALDI-TOF mass spectrometry in clinical diagnostic microbiology. **The Journal of Infection in Developing Countries**, v. 8, n. 09, p. 1081-1088, 2014.

DE SOUTO, M. C. P. *et al*. Técnicas de aprendizado de máquina para problemas de biologia molecular. **Sociedade Brasileira de Computação**, v. 1, n. 2, 2003.

ESTER, M. *et al*. A density-based algorithm for discovering clusters in large spatial databases with noise. In: **kdd**. 1996. p. 226-231.

ESTIGARRIBIA, D. A. C. **Análise de biomarcadores proteicos para estudos de identificação, resistência e variabilidade em Staphylococcus spp. resistentes à meticilina (MRS) e Enterococcus resistentes à vancomicina (VRE)**. 2017. Dissertação (Pós-Graduação em Biotecnologia e Biociências) - Universidade Federal de Santa Catarina, Florianópolis, 2017.

GANS, J.; WOLINSKY, M.; DUNBAR, J. Computational improvements reveal great bacterial diversity and high metal toxicity in soil. **Science**, v. 309, n. 5739, p. 1387-1390, 2005.

GARCÍA, P. *et al*. Identificación bacteriana basada en el espectro de masas de proteínas: Una nueva mirada a la microbiología del siglo XXI. **Revista chilena de infectología**, v. 29, n. 3, p. 263-272, 2012.

HOTTA, Y. *et al.* Classification of genus pseudomonas by MALDI-TOF MS based on ribosomal protein coding in S10− spc− alpha operon at strain level. **Journal of proteome research**, v. 9, n. 12, p. 6722-6728, 2010.

HOU, T. Y.; CHIANG-NI, C.; TENG, S. H.. Current status of MALDI-TOF mass spectrometry in clinical microbiology. **Journal of food and drug analysis**, v. 27, n. 2, p. 404-414, 2019.

JÚNIOR, P. S. P. C. *et al*. Endophytic bacteria of garlic roots promote growth of micropropagated meristems. **Microbiological research**, v. 241, p. 126585, 2020.

JURINKE, C.; OETH, P.; VAN DEN BOOM, D. MALDI-TOF mass spectrometry. **Molecular biotechnology**, v. 26, n. 2, p. 147-163, 2004.

KUMAR, A. *et al*. Plant growth-promoting bacteria: biological tools for the mitigation of salinity stress in plants. **Frontiers in Microbiology**, v. 11, p. 1216, 2020.

LAUBER, M. A.; RUNNING, W. E.; REILLY, J. P. B. subtilis ribosomal proteins: structural homology and post-translational modifications. **Journal of proteome research**, v. 8, n. 9, p. 4193-4206, 2009.

LEGARRAGA, P. *et al*. Impacto de la espectrometría de masas por MALDI-TOF MS en la identificación rápida de bacterias aeróbicas y anaeróbicas de importancia clínica. **Revista chilena de infectología**, v. 30, n. 2, p. 140-146, 2013.

LORENA, A. C.; GAMA, J.; FACELI, K. Inteligência Artificial: Uma abordagem de aprendizado de máquina. [S.l.]: Grupo Gen-LTC, 2000.

MADHULATHA, T. S. An overview on clustering methods. arXiv preprint arXiv:1205.1117,2012.

MAJEED, A.; MUHAMMAD, Z.; AHMAD, H.. Plant growth promoting bacteria: role in soil improvement, abiotic and biotic stress management of crops. **Plant cell reports**, v. 37, n. 12, p. 1599-1609, 2018.

MAYER, B.; HOLTRUP, S.; GRAUMANN, P. L. A Machine Learning-Empowered Workflow to Discriminate Bacillus subtilis Motility Phenotypes. **BioMedInformatics**, v. 2, n. 4, p. 565-579, 2022.

MENDES, R.; GARBEVA, P.; RAAIJMAKERS, J. M. The rhizosphere microbiome: significance of plant beneficial, plant pathogenic, and human pathogenic microorganisms. **FEMS microbiology reviews**, v. 37, n. 5 , p. 634-663, 2013.

MUTHURI, C. *et al*. Isolation an d identification of endophytic bacteria of bananas (Musa spp.) in Kenya and thei r potential as biofertilizers for sustainable banana production. **Afr. J. Microbiol. Res**, v. 6, p. 6414-6422, 2012.

NASCIMENTO, R. S. **Identificação de proteínas ribossomais em espectro de massa do tipo Maldi-Tof.** 2019. Dissertação (Mestrado em Computação Aplicada) - Universidade Estadual de Ponta Grossa, Ponta Grossa, 2019.

PEDREGOSA, F. *et al*. Scikit-learn: Machine learning in python. **Journal of machine learning research**, v. 12, n. Oct, p. 2825–2830, 2011.

RIBEIRO, M. C. M. **Genética molecular**. CED/LANTEC, 2009.

SCHUBERT, E. *et al*. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. **ACM Transactions on Database Systems (TODS)**, v. 42, n. 3, p. 1-21, 2017.

SILVA, R. R. da. **Classificação bacteriana baseada em proteínas ribossomais oriundas de dados genômicos**. 2021. Dissertação (Mestrado em Computação Aplicada) - Universidade Estadual de Ponta Grossa, Ponta Grossa, 2021.

SINGH, B. K. *et al*. Emerging microbiome technologies for sustainable increase in farm productivity and environmental security. **Microbiology Australia**, v. 39, n. 1, p. 17-23, 2018.

TAMURA, H.; HOTTA, Y.; SATO, H. Novel accurate bacterial discrimination by MALDI-time-of-flight MS based on ribosomal proteins coding in S10-spc-alpha operon at strain level S10-GERMS. **Journal of The American Society for Mass Spectrometry**, v. 24, n. 8, p. 1185-1193, 2013.

TERAMOTO, K. *et al*. Phylogenetic classification of pseudomonas putida strains by maldi-ms using ribosomal subunit proteins as biomarkers. **Analytical chemistry**, ACS Publications, v. 79, n. 22, p. 8712–8719, 2007.

TOMACHEWSKI, D. **Utilização de aprendizado de máquina para classificação de bactérias através de proteínas ribossomais**. 2017, 72f. Dissertação (Mestrado em Computação Aplicada), Universidade Estadual de Ponta Grossa, Ponta Grossa, 2017.

TOMACHEWSKI, D. *et al*. Ribopeaks: a web tool for bacterial classification through m/z data from ribosomal proteins. Bioinformatics, **Oxford University Press**, v. 34, n. 17, p. 3058–3060, 2018.

ZAUTNER, A. E. *et al*. Mass Spectrometry-based PhyloProteomics (MSPP): A novel microbial typing Method. **Scientific reports**, v. 5, n. 1, p. 13431, 2015.

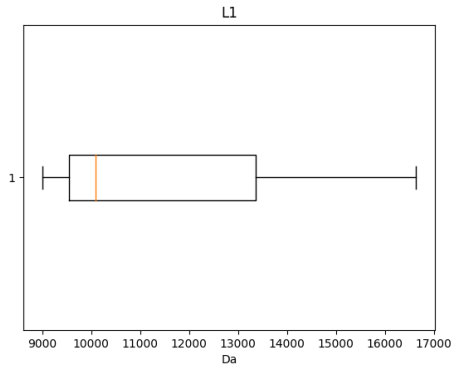
ZIEGLER, D. et al. Ribosomal protein biomarkers provide root nodule bacterial identification by MALDI-TOF MS. **Applied microbiology and biotechnology**, v. 99, n. 13, p. 5547-5562, 2015.

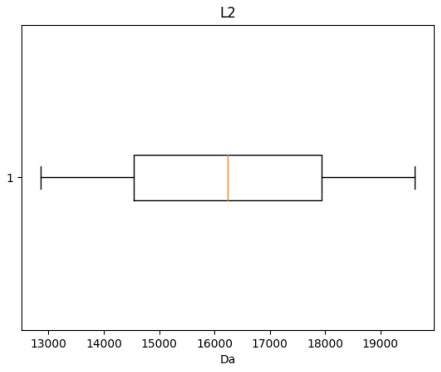
APÊNDICE A - VALORES DOS QUARTIS DAS PROTEÍNAS RIBOSSOMAIS

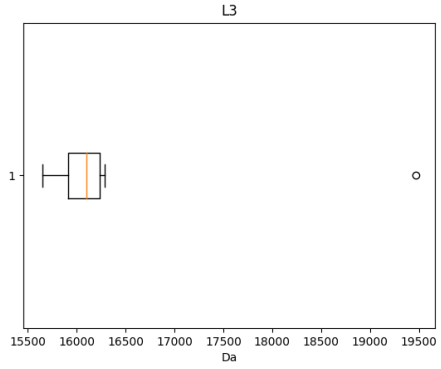
Proteína	Quantidade	Outliers	Q1	Q2	Q3	IIQ
L1	3	0	9542.6590	10088.8263	13365.6938	3823.0347
L10	4031	7	17693.4638	18087.8744	18609.1565	915.6927
L11	2158	169	14847.2687	15005.1015	15237.6972	390.4284
L13	3868	97	16023.3930	16311.9383	16735.4986	712.1056
L14	2804	100	13236.4391	13331.7828	13417.8415	181.4024
L15	4074	200	15230.9255	15626.6185	16199.9336	969.0081
L16	3308	65	15506.2413	15715.1462	15984.0270	477.7857
L17	3529	0	14064.1089	14886.4480	16689.7085	2625.5996
L18	3871	199	12719.1690	12993.8733	13345.2340	626.0651
L19	3822	319	13150.8656	13452.7397	14123.3834	972.5179
L2	2	0	14544.1650	16237.8136	17931.4623	3387.2974
L20	3509	140	13264.5177	13487.6101	13869.2681	604.7504
L21	3526	554	11222.1599	11466.2476	11740.7662	518.6063
L22	3394	105	12261.6962	12895.8712	14114.0276	1852.3314
L23	3608	293	10774.5610	11019.1398	11276.2707	501.7097
L24	3753	475	11117.8648	11357.2064	11822.0064	704.1416
L25	667	158	10535.2186	10663.4366	11874.2811	1339.0625
L27	3494	170	8969.2861	9232.5301	9704.0712	734.7851
L28	3382	11	6980.5287	8742.2693	9013.5680	2033.0393
L29	3268	218	7373.3542	7739.9596	8268.9774	895.6232
L3	7	1	15913.2294	16101.4968	16234.516	321.2865
L30	2864	203	6485.6551	6631.8201	6849.2812	363.6261
L31	4171	51	7979.0258	8980.1735	9744.6422	1765.6163
L32	3252	71	6375.1045	6621.5014	6972.3016	597.1972
L33	2947	179	5989.0672	6241.2803	6461.3852	472.3180
L34	2300	326	5161.1468	5274.3263	5474.7258	313.5791
L35	3125	86	7187.7239	7314.9122	7443.9554	256.2315
L36	1920	92	4352.0936	4451.4573	4798.0646	445.9710
L4	5	1	5700.7603	5729.7114	8417.4897	2716.7294

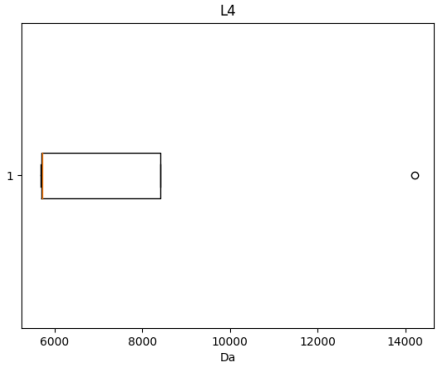
L5	526	16	19812.1244	19898.6371	19956.8654	144.7410
L6	4144	8	18923.0385	19197.9136	19419.7541	496.7156
L7A	41	0	8654.0392	8753.1718	10980.9963	2326.9571
L7AE	544	3	8712.2150	9562.5920	11465.8806	2753.6656
L7L12	4033	118	12434.1781	12654.4335	13047.0478	612.8697
L9	3947	321	15847.2548	16222.9684	16570.9759	723.7211
S1	6	1	13166.4612	13973.1683	14198.7854	1032.3242
S10	2150	106	11454.3261	11604.7091	11781.9160	327.5899
S11	2853	112	13639.8480	13791.2047	14028.9794	389.1314
S12	1803	1	13653.9388	13843.3599	15046.1090	1392.1702
S13	3551	100	13405.6427	13666.6745	13937.2313	531.5886
S14	3614	1	7197.5835	11258.3238	11484.2166	4286.6331
S15	3592	107	10087.5379	10232.9304	10391.1189	303.581
S16	3750	0	9305.0065	10224.9248	15299.9509	5994.9444
S17	3525	223	9662.4654	10014.6310	10426.0065	763.5411
S18	4743	0	10046.8088	16482.6518	17333.4756	7286.6668
S19	2807	63	10132.1942	10315.8603	10463.9772	331.7831
S2	1	0	14641.0618	14641.0618	14641.0618	0
S20	3849	389	9316.7880	9517.1627	9765.3652	448.5772
S21	2117	110	7488.8571	8109.6963	8459.8687	971.0116
S22	12	1	5194.3869	5250.9951	5272.2777	77.8908
S3	6	0	9220.6284	9511.9780	9575.2068	354.5784
S4	6	0	13295.8973	15731.0327	18125.7542	4829.8570
S5	2408	204	17401.9799	17701.0308	18092.1908	690.2109
S6	3449	0	11403.2301	13841.9722	15514.7747	4111.5446
S7	3417	178	17493.9944	17658.6023	17832.7188	338.7244
S8	3665	93	14111.5525	14393.5479	14638.8340	527.2815
S9	3800	9	14361.7772	14575.9121	17105.3846	2743.6074
THX	208	7	4210.3319	4774.2030	5756.8581	1546.5263
YCF65	59	0	11439.0904	12207.7666	17394.0696	5954.9792

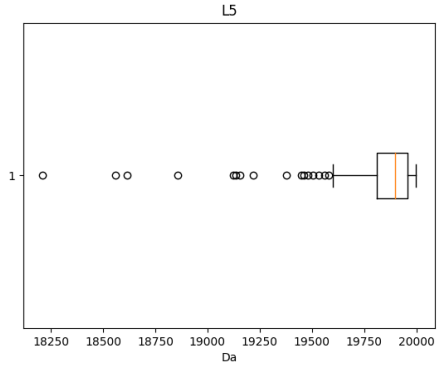
APÊNDICE B - BOXPLOT DAS PROTEÍNAS RIBOSSOMAIS

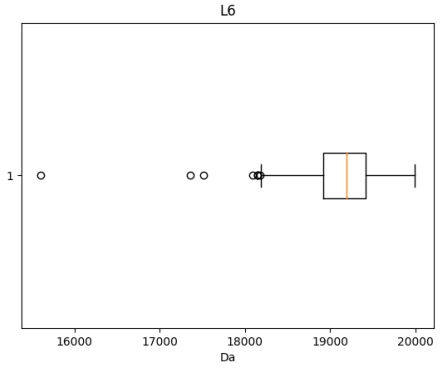


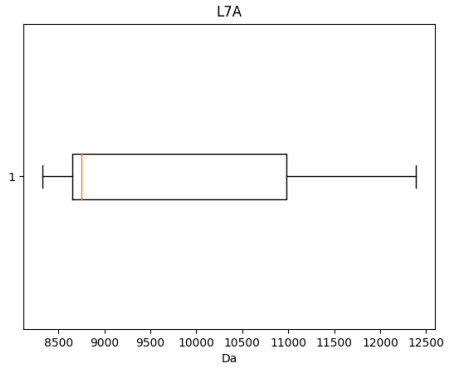


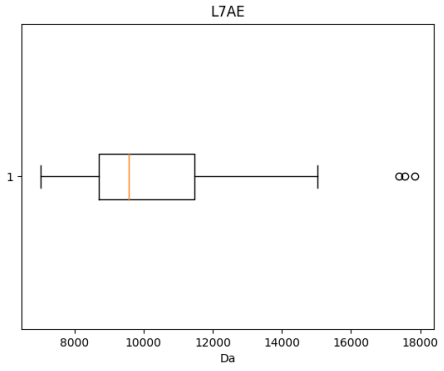


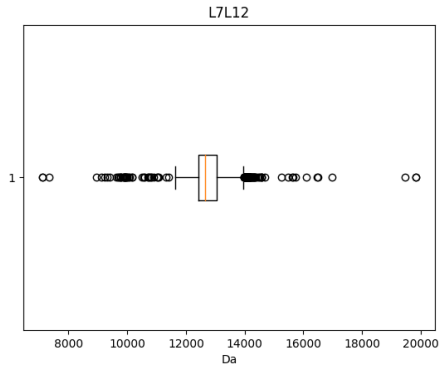


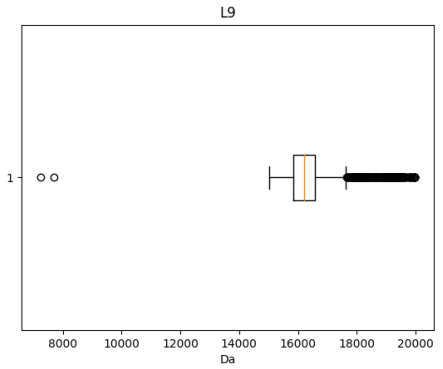


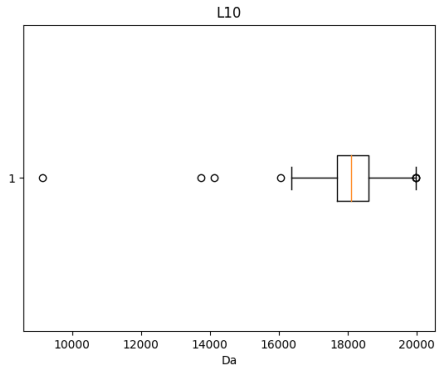


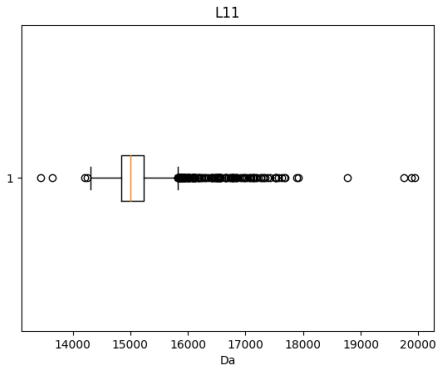


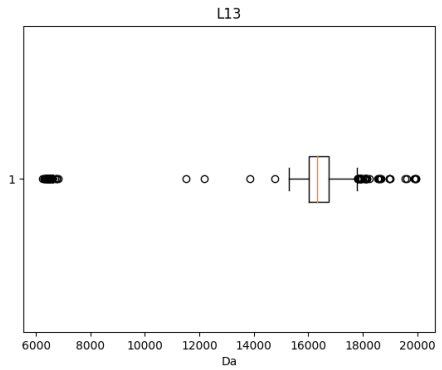


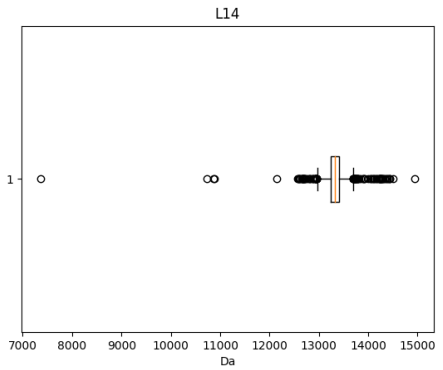

L7L12
1
8000
10000
12000
14000
16000
18000
20000
Da

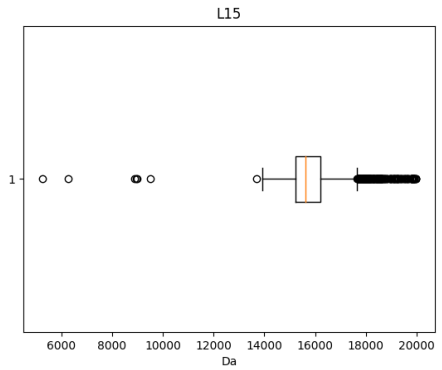

L9
1
8000
10000
12000
14000
16000
18000
20000
Da

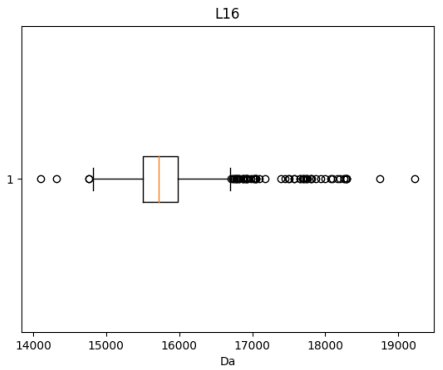

L10
1
10000
12000
14000
16000
18000
20000
Da

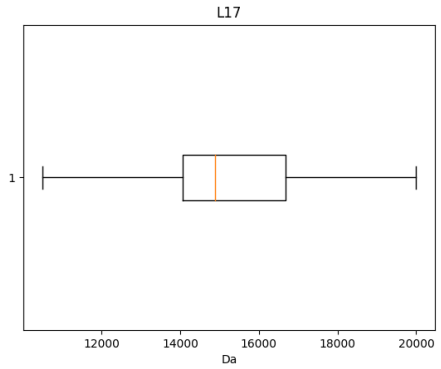

L11
1
14000
15000
16000
17000
18000
19000
20000
Da

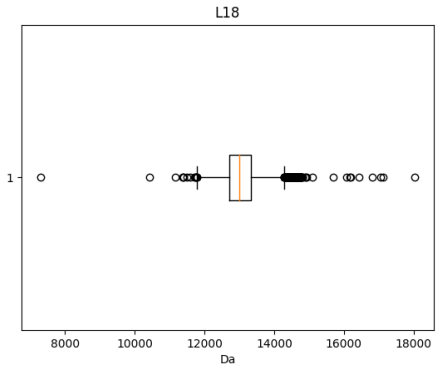

L13
1
6000
8000
10000
12000
14000
16000
18000
20000
Da

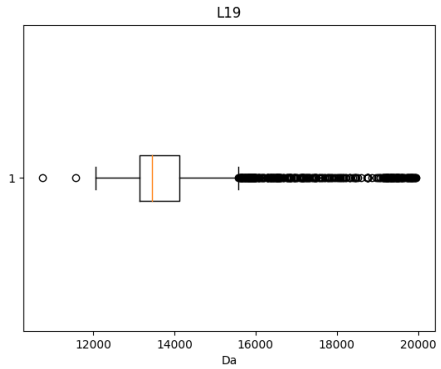

L14
1
7000
8000
9000
10000
11000
12000
13000
14000
15000
Da

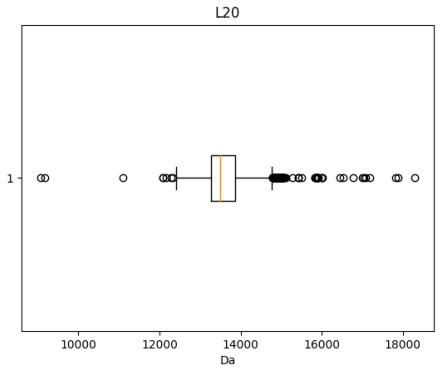

L15
1
6000
8000
10000
12000
14000
16000
18000
20000
Da

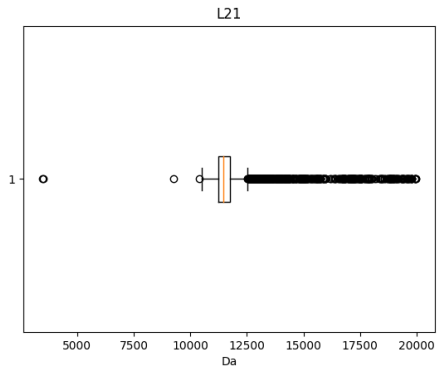

L16
1
14000
15000
16000
17000
18000
19000
Da

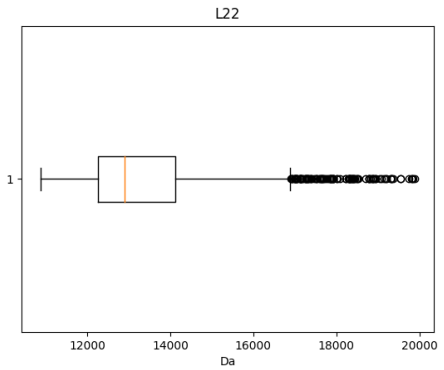

L17
1
12000
14000
16000
18000
20000
Da

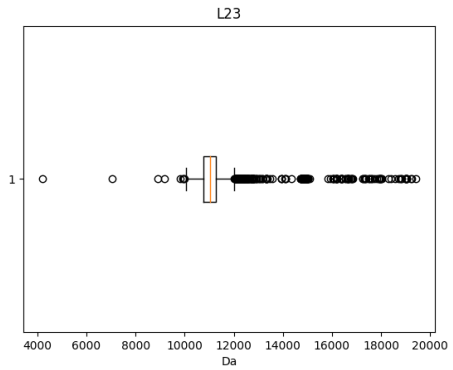

L18
1
8000
10000
12000
14000
16000
18000
Da

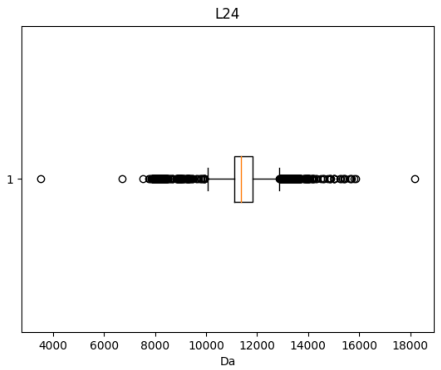

L19
1
12000
14000
16000
18000
20000
Da

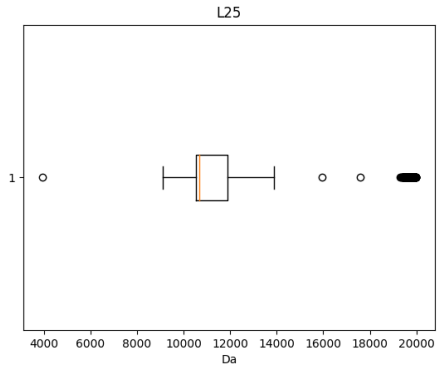

L20
1
10000
12000
14000
16000
18000
Da

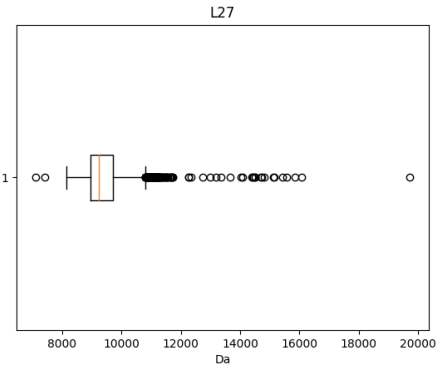

L21
1
5000
7500
10000
12500
15000
17500
20000
Da

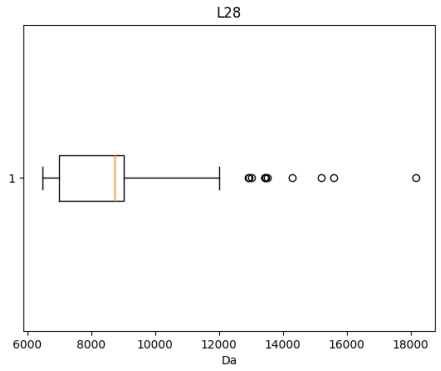

L22
1
12000
14000
16000
18000
20000
Da

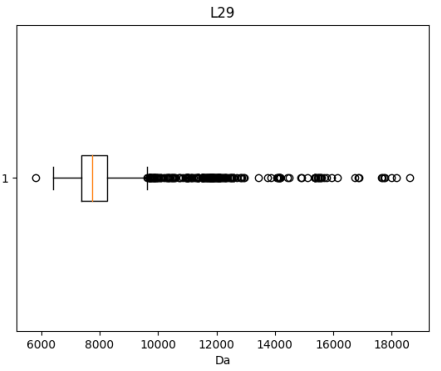

L23
1
4000
6000
8000
10000
12000
14000
16000
18000
20000
Da

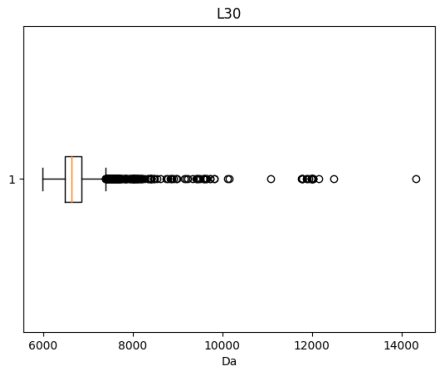

L24
1
4000
6000
8000
10000
12000
14000
16000
18000
Da

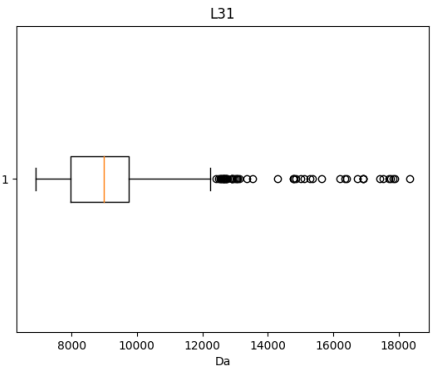

L25
1
4000 6000 8000 10000 12000 14000 16000 18000 20000
Da

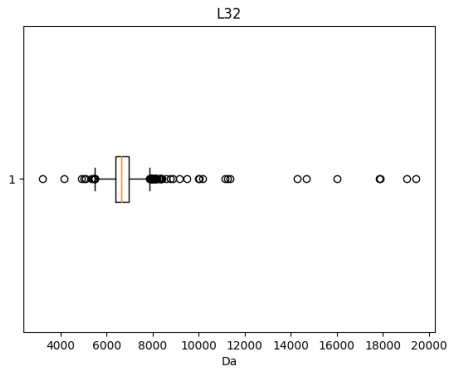

L27
1
8000 10000 12000 14000 16000 18000 20000
Da

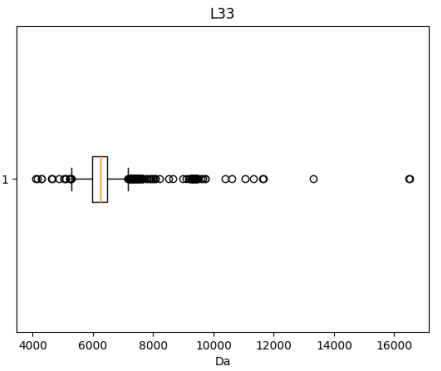

L28
1
6000 8000 10000 12000 14000 16000 18000
Da

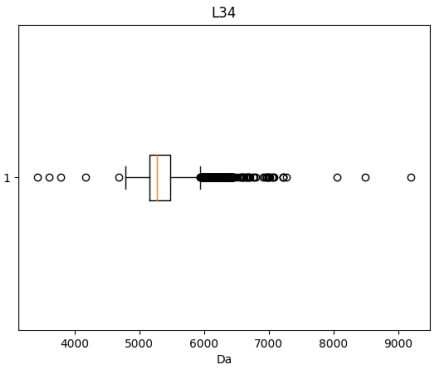

L29
1
6000 8000 10000 12000 14000 16000 18000
Da

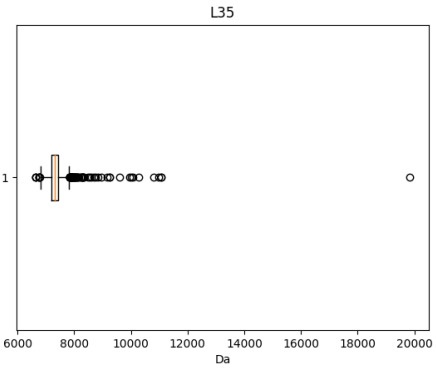

L30
1
6000 8000 10000 12000 14000
Da

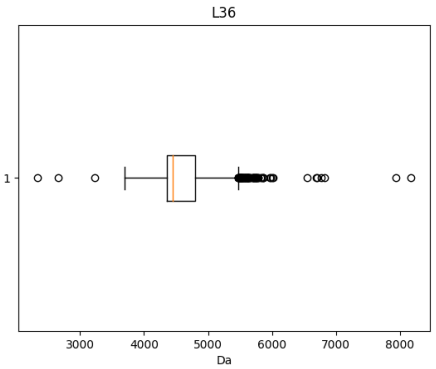

L31
1
8000 10000 12000 14000 16000 18000
Da

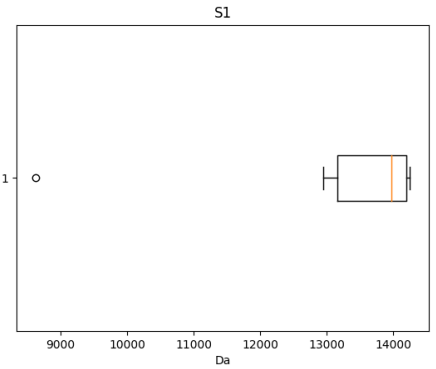

L32
1
4000 6000 8000 10000 12000 14000 16000 18000 20000
Da

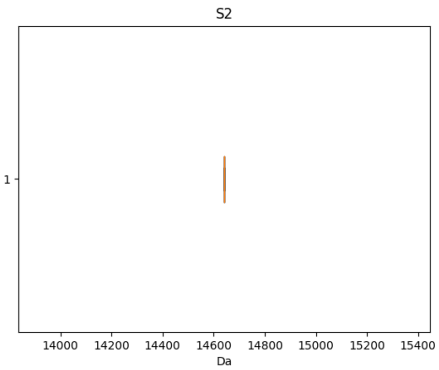

L33
1
4000 6000 8000 10000 12000 14000 16000
Da

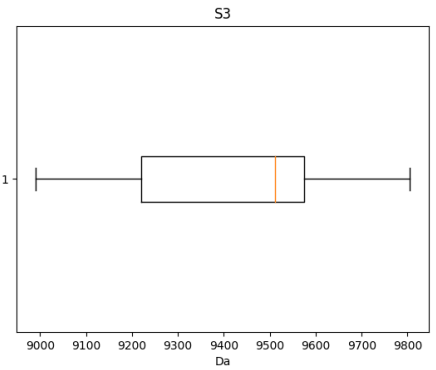

L34
1
4000
5000
6000
7000
8000
9000
Da

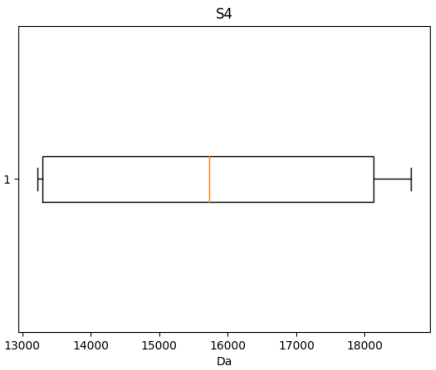

L35
1
6000
8000
10000
12000
14000
16000
18000
20000
Da

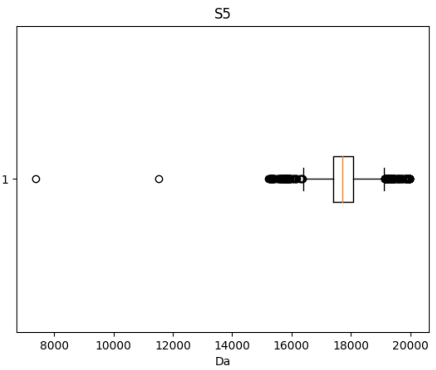

L36
1
3000
4000
5000
6000
7000
8000
Da

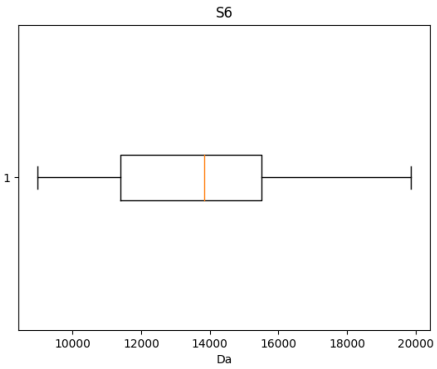

S1
1
9000
10000
11000
12000
13000
14000
Da

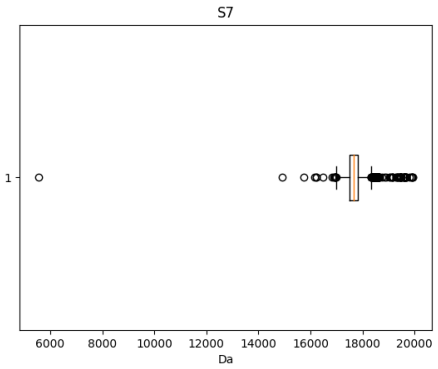

S2
1
14000
14200
14400
14600
14800
15000
15200
15400
Da

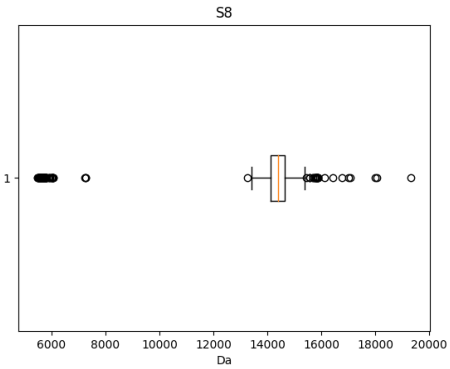

S3
1
9000
9100
9200
9300
9400
9500
9600
9700
9800
Da

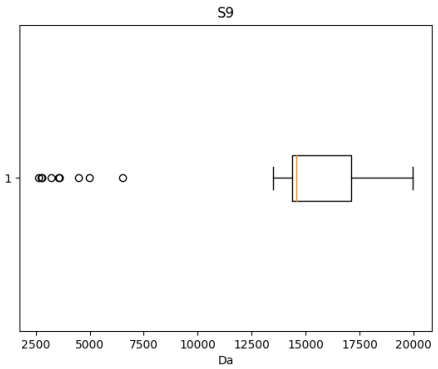

S4
1
13000
14000
15000
16000
17000
18000
Da

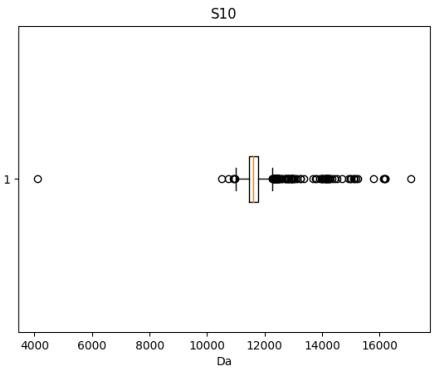

S5
1
8000
10000
12000
14000
16000
18000
20000
Da

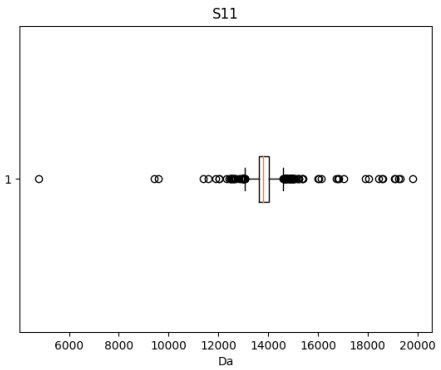

S6
1
10000
12000
14000
16000
18000
20000
Da

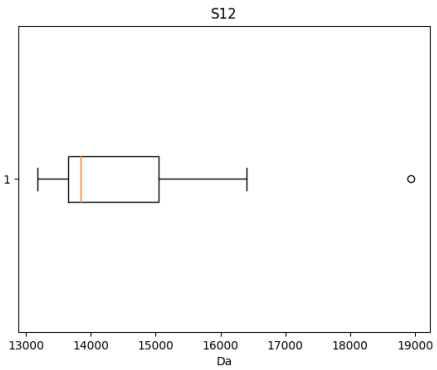

S7
1
6000
8000
10000
12000
14000
16000
18000
20000
Da

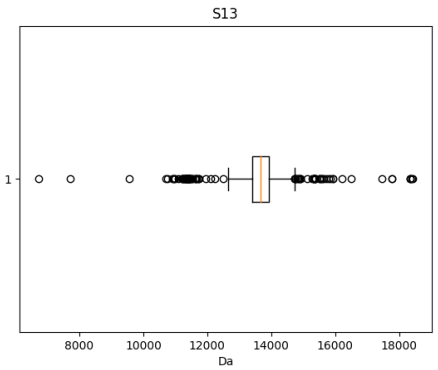

S8
1
6000
8000
10000
12000
14000
16000
18000
20000
Da

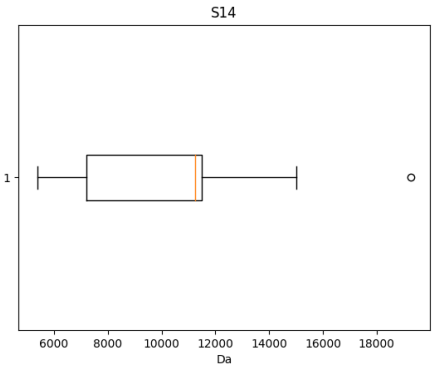

S9
1
2500
5000
7500
10000
12500
15000
17500
20000
Da

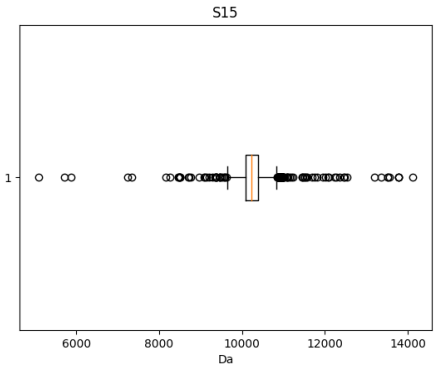

S10
1
4000
6000
8000
10000
12000
14000
16000
Da

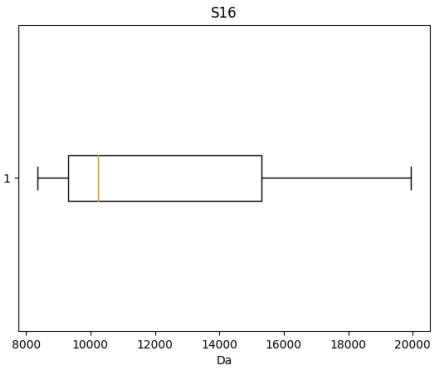

S11
1
6000
8000
10000
12000
14000
16000
18000
20000
Da

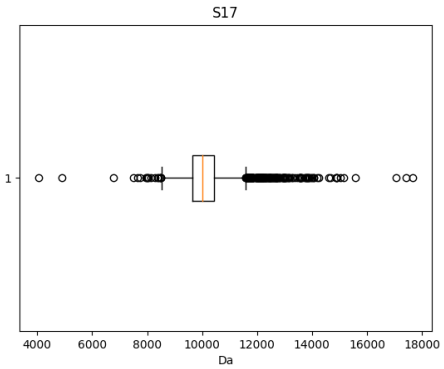

S12
1
13000
14000
15000
16000
17000
18000
19000
Da

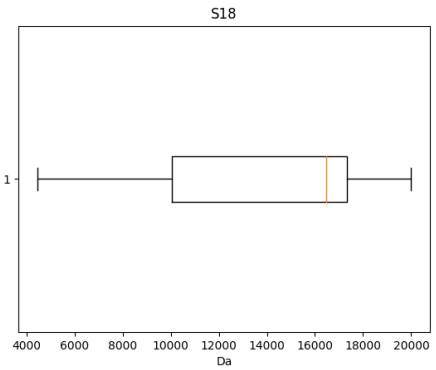

S13
1
8000
10000
12000
14000
16000
18000
Da

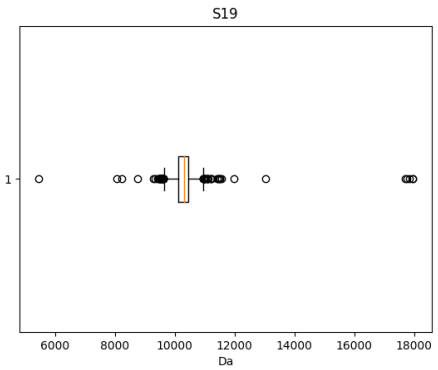

S14
1
6000
8000
10000
12000
14000
16000
18000
Da

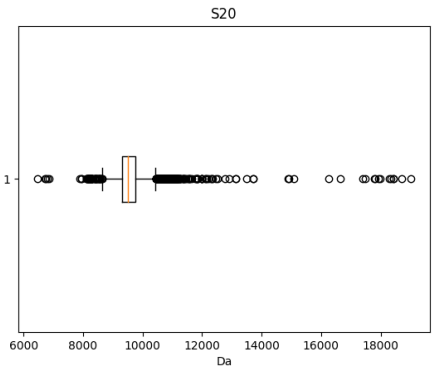

S15
1
6000
8000
10000
12000
14000
Da

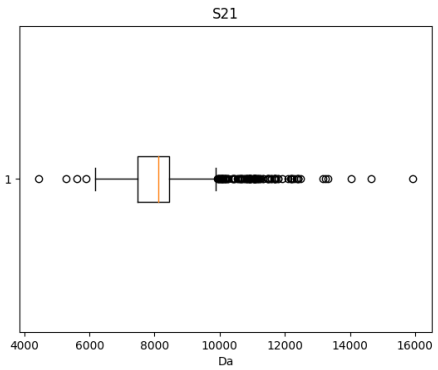

S16
1
8000
10000
12000
14000
16000
18000
20000
Da

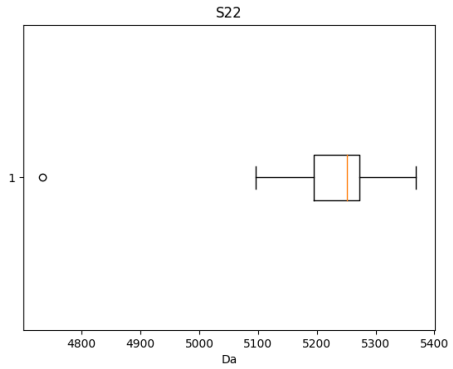

S17
1
4000
6000
8000
10000
12000
14000
16000
18000
Da

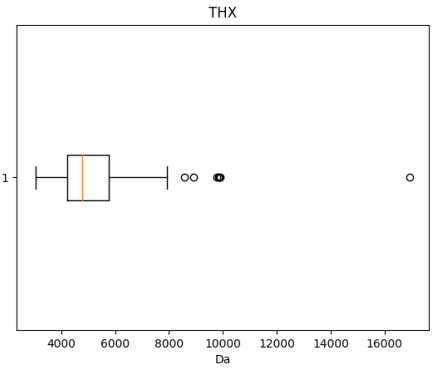

S18
1
4000
6000
8000
10000
12000
14000
16000
18000
20000
Da

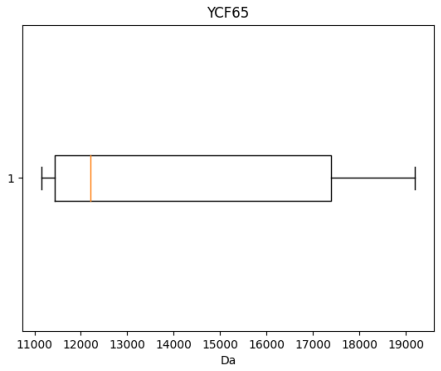

S19
1
6000
8000
10000
12000
14000
16000
18000
Da


S20
1
6000
8000
10000
12000
14000
16000
18000
Da


S21
1
4000
6000
8000
10000
12000
14000
16000
Da


S22
1
4800
4900
5000
5100
5200
5300
5400
Da


THX
1
4000
6000
8000
10000
12000
14000
16000
Da


YCF65
1
11000
12000
13000
14000
15000
16000
17000
18000
19000
Da

APÊNDICE C - PROTEÍNAS COM VALOR DE M/Z EM COMUM

Proteína 1	Proteína 2	Pico Comum (Da)
L33	L30	6369.5538
L33	L30	6447.6519
L33	L30	6475.6653
S21	L28	6614.8095
L33	L29	7166.5718
L35	L31	7576.8674
L7A	L7AE	8330.8524
L7A	L7AE	8735.2808
S18	L31	8901.2665
S16	L28	8947.3603
S20	L31	8995.2282
S20	L28	8999.5664
L27	L25	9111.5504
S20	L28	9303.9715
S20	S18	9606.2575
L31	L27	9628.1258
S20	S18	9728.3836
S19	S17	9871.4095
S20	S17	9893.5787
S19	S16	9994.6101
S19	S16	10124.7621
S16	S15	10131.7383
S15	L27	10158.6625
S19	S17	10183.9741
S19	S14	10280.1281
S19	S15	10286.8668
S19	S16	10324.9188
S19	S15	10403.1298
S15	L23	10412.1189
S15	L27	10456.1676
S19	L28	10468.1066
S17	L23	10542.1943
S15	L23	10566.3302
L27	L23	10757.4233
L23	L21	10969.6145
L7A	L7AE	10980.9963
S10	L21	11666.7476
L24	L23	11775.7307
S10	L22	11802.7830
L23	L22	11836.8639
S13	L14	13015.1400
L19	L18	13019.0301
L19	L14	13133.3513
L19	L14	13171.3911
S13	L19	13239.4789
L20	L14	13275.5813
L20	L14	13348.7316
L20	L19	13348.8817
L19	L17	13380.6613
L18	L14	13389.6599
S13	S11	13471.6257
S12	L19	13481.8885
S13	S12	13536.7700
S13	S12	13562.8513
L19	L18	13610.0343
S12	L14	13660.9343
S12	L14	13674.9612
L22	L19	13757.9263
S13	L22	13822.0265
S8	S13	13956.3152
S8	S11	14093.1711
S6	S11	14261.3206
S8	L19	14504.7750
S9	L17	14511.7380
S9	L11	14847.2576
S12	L11	15132.5889
L17	L15	15558.8952
L9	L17	15563.0083
L9	L13	15648.1717
L15	L13	15862.3713
L15	L13	15892.3976
S16	L16	16095.9490

S16	L23	11010.9199	L9	L13	16304.8802
L24	L21	11149.0358	L17	L13	16389.1159
L24	L23	11156.8856	L15	L13	16417.9319
L7A	L7AE	11166.1535	L9	L15	16554.2069
L24	L21	11170.0225	S5	L13	17049.7087
L23	L21	11174.8155	S7	S16	17115.8251
L7A	L7AE	11220.1693	S7	S5	17366.0446
S14	L21	11305.3736	S7	S5	17369.1593
S14	S10	11319.1670	S7	S5	17443.1954
S10	L24	11402.2495	S7	L10	17675.5562
S6	L21	11473.2997	S7	L10	17739.3481
S10	L21	11562.4713	S5	L10	18162.3427
S10	L21	11611.5283	S16	L6	18757.5957
S16	L24	11649.5624	L6	L10	19085.2650

APÊNDICE D - QUANTIDADE DE PROTEÍNAS POR ORGANISMO

	Quantidade de Picos															
		Base filtrada de acordo com cada modelo DBSCAN														
Organismo	SpectraBank	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5
Acinetobacter baumannii ATCC 15308	13	7	8	10	10	11	11	11	11	11	12	12	12	12	12	12
Aeromonas hydrophila ATCC 7966	26	9	12	13	15	15	16	16	16	16	16	17	17	17	17	17
Bacillus amyloliquefaciens ATCC 23842	30	8	10	11	14	15	15	15	15	16	16	17	17	17	17	17
Bacillus cereus ATCC 14579	17	4	8	9	11	11	11	11	11	11	11	11	11	11	11	11
Bacillus cereus ATCC 14893	18	2	3	4	7	7	7	7	7	7	7	7	7	7	7	7
Bacillus cereus ATCC 9634	15	2	3	6	7	8	8	8	8	10	10	10	10	10	10	10
Bacillus licheniformis 1Pesc	15	4	4	4	5	5	5	5	6	6	6	6	6	6	6	6
Bacillus licheniformis ATCC 14580	16	6	6	6	6	6	6	6	6	6	6	6	6	6	8	8
Bacillus licheniformis ATCC 27811	12	5	5	5	5	6	6	6	6	7	7	8	8	8	8	8
Bacillus megaterium ATCC 14581	17	4	5	8	9	10	11	11	11	11	11	11	11	11	11	11
Bacillus megaterium ATCC 25848	17	5	9	11	12	12	12	12	12	12	12	12	12	12	12	12
Bacillus megaterium RPesc	18	6	8	11	11	11	12	12	12	12	12	12	12	12	12	13
Bacillus pumilus ATCC 14884	15	3	7	10	10	10	10	10	10	10	10	10	10	10	10	10
Bacillus pumilus ATCC 7061	14	4	5	7	9	9	9	9	9	9	9	9	9	10	10	10
Bacillus sp. (amyloliquefaciens polyfermenticus subtilis) Proc5	27	9	11	15	16	18	18	18	18	18	18	19	19	19	19	19
Bacillus sp. (amyloliquefaciens polyfermenticus subtilis) Proc7a	13	4	7	8	8	8	8	8	8	8	8	8	9	9	9	9
Bacillus sp. (amyloliquefaciens polyfermenticus subtilis) ProcB5	21	7	10	11	14	15	15	15	15	16	16	16	16	16	16	16
Bacillus sp. (cereus thuringiensis) 20MC1	30	0	5	7	10	11	12	12	13	13	13	13	13	13	13	13
Bacillus sp. (cereus thuringiensis) Proc3T9	19	2	4	7	8	8	8	8	8	8	8	8	8	8	8	8
Bacillus sp. (cereus thuringiensis) ProcB2	24	2	3	4	5	6	6	6	6	6	6	6	6	6	6	6
Bacillus sp. (cereus thuringiensis) Seab10	24	5	5	7	8	11	12	12	12	12	12	12	12	12	12	12
Bacillus sp. (cereus thuringiensis) Seab21	19	2	4	6	8	8	8	8	8	8	8	8	8	8	8	8
Bacillus subtilis ATCC 11774	23	9	10	11	14	14	14	14	14	14	14	14	14	14	14	14
Bacillus subtilis ATCC 6051	30	10	10	13	14	14	14	14	14	14	15	15	15	15	15	15
Bacillus subtilis ATCC 6633	21	6	8	11	12	12	12	13	13	13	13	13	13	13	13	13
Bacillus subtilis ATCC 9524	28	7	7	11	12	13	13	13	13	14	14	14	14	14	14	14
Bacillus subtilis Proc6a	30	6	10	15	17	17	17	17	17	17	17	17	17	17	17	17
Bacillus subtilis Proc6b	34	6	9	11	12	12	12	12	12	13	13	13	13	13	13	13
Bacillus subtilis Proc6c	23	3	6	7	9	12	12	12	12	12	12	12	12	12	12	12
Bacillus subtilis Proc721	30	6	7	9	11	12	12	12	12	13	13	13	14	14	14	14
Bacillus subtilis ProcB6c	27	5	10	12	15	16	16	16	16	16	17	17	17	17	17	17
Bacillus thuringiensis ATCC 10792	15	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Bacillus thuringiensis ATCC 33679	18	5	5	5	7	7	7	7	7	7	7	7	7	7	7	7
Bacillus thuringiensis ATCC 35866	16	2	3	4	4	5	5	5	6	6	6	6	6	6	6	6

Carnobacterium divergens ATCC 35677	12	2	3	3	3	3	3	3	3	4	4	5	5	6	6	6
Carnobacterium gallinarum ATCC 49517	12	4	6	6	6	6	6	6	6	7	7	7	7	7	7	7
Carnobacterium maltaromaticum ATCC 27865	17	7	8	10	11	11	11	11	11	11	12	12	12	12	12	12
Carnobacterium maltaromaticum ATCC 35586	15	5	6	7	7	7	7	8	8	9	9	9	9	9	9	9
Carnobacterium maltaromaticum Proc3T4	13	3	3	4	5	6	6	6	6	7	8	8	8	8	8	8
Carnobacterium maltaromaticum Proc4T4	13	3	4	4	5	6	6	6	6	7	7	7	7	7	7	7
Citrobacter freundii ATCC 8090	31	9	14	17	18	19	20	21	21	21	22	22	22	22	22	22
Clostridium botulinum ATCC 19397	27	6	7	12	13	13	13	13	14	14	14	14	14	15	15	15
Clostridium perfringens ATCC 10543	24	7	10	10	11	12	12	12	12	12	12	12	12	12	12	12
Enterobacter aerogenes ATCC 13048	25	10	14	14	15	15	15	15	15	15	15	15	15	15	16	16
Enterobacter cloacae ATCC 13047	24	11	13	14	15	15	15	15	15	15	15	15	15	15	15	15
Enterobacter hormaechei 10MC1	26	6	9	14	14	14	14	14	14	14	15	15	15	15	15	15
Enterobacter sakazakii ATCC 29544	24	2	5	9	10	12	12	12	12	12	12	12	12	12	12	12
Escherichia coli NCTC 50271	34	12	17	20	22	22	22	22	22	22	23	23	23	23	23	23
Escherichia coli NCTC 50365	31	10	14	16	20	20	21	21	21	21	21	21	21	21	21	21
Hafnia alvei ATCC 9760	25	5	8	12	13	14	14	14	14	14	14	14	15	15	15	15
Klebsiella oxytoca ATCC 13182	34	13	16	18	22	22	23	23	23	23	23	23	23	23	24	24
Klebsiella pneumoniae ATCC 10031	18	5	7	8	9	9	9	9	10	10	10	10	10	10	10	10
Listeria innocua ATCC 33090	30	7	13	17	21	21	21	21	21	21	21	21	21	21	21	21
Listeria ivanovii ATCC 19119	35	11	14	19	19	20	20	20	20	20	20	20	20	20	20	20
Listeria monocytogenes CECT 4032	31	10	14	18	19	20	20	20	20	20	20	20	20	20	20	20
Listeria seeligeri ATCC 35967	31	15	16	17	18	20	20	20	20	20	20	20	20	20	20	20
Listeria welshimeri ATCC 35897	29	7	14	17	17	18	18	18	18	18	18	18	18	18	19	19
Morganella morganii ATCC 8076	34	11	17	19	21	22	22	22	22	22	23	24	24	24	24	24
Morganella morganii BM 65	26	7	8	13	14	14	14	16	17	17	17	17	17	17	17	17
Pantoea agglomerans ATCC 27155	28	7	14	16	16	17	19	19	19	19	19	19	19	19	19	20
Photobacterium damselae ATCC 33539	30	9	13	15	19	19	19	19	19	19	19	19	19	19	19	19
Photobacterium phosphoreum CECT 4172	34	14	18	20	22	22	22	22	22	22	23	24	24	24	24	24
Proteus mirabilis ATCC 14153	25	6	9	12	14	14	14	14	14	14	14	14	14	15	16	16
Proteus penneri ATCC 33519	23	5	5	6	8	10	10	10	10	11	11	11	11	12	12	12
Proteus vulgaris ATCC 9484	23	4	7	10	12	12	12	12	12	12	12	12	13	14	14	14
Proteus vulgaris sard1	22	4	6	7	9	10	10	11	11	11	11	12	12	12	12	12
Proteus vulgaris sard2	20	5	7	8	10	11	12	12	12	12	12	12	13	13	13	13
Proteus vulgaris sard3	23	4	6	8	11	12	12	12	12	12	12	12	12	13	13	13
Proteus vulgaris sard4	24	6	10	14	15	15	16	16	16	16	16	16	17	17	17	17
Providencia rettgeri ATCC 29944	27	8	13	15	16	16	17	17	17	18	18	18	18	18	18	18
Providencia stuartii ATCC 29914	20	8	9	9	10	10	10	10	10	10	10	10	10	10	11	11
Pseudomonas fluorescens ATCC 13525	28	10	13	16	17	19	19	19	19	19	19	19	19	19	19	19
Pseudomonas fluorescens ATCC 17397	24	9	13	13	14	15	15	15	15	15	15	15	15	15	15	16
Pseudomonas fluorescens Turb28	34	6	11	16	16	16	16	16	16	17	17	17	17	18	19	19
Pseudomonas fluorescens Turb46	27	7	12	13	14	14	15	15	15	15	15	15	15	15	16	16

Pseudomonas fluorescens Turb52	20	3	6	7	10	10	10	10	10	10	10	11	11	11	11	11
Pseudomonas fluorescens Turb64	22	7	8	9	11	12	12	13	13	13	13	13	13	13	13	13
Pseudomonas fragi ATCC 4973	25	9	12	15	17	18	18	18	18	18	18	18	18	18	18	18
Pseudomonas fragi Seab03	22	5	12	13	13	14	15	15	15	15	15	15	15	15	15	15
Pseudomonas fragi Seab22	21	8	11	13	13	13	14	14	14	14	14	14	14	14	14	14
Pseudomonas fragi Seab23	22	9	11	13	13	13	13	14	14	14	14	14	14	14	14	15
Pseudomonas fragi Turb32	24	8	11	13	14	15	15	15	15	15	15	15	15	15	15	15
Pseudomonas fragi Turb43	25	7	8	10	12	12	14	14	14	14	14	14	14	14	15	15
Pseudomonas fragi Turb47	27	5	7	10	14	15	15	15	15	15	15	16	16	16	16	16
Pseudomonas putida ATCC 12633	28	13	14	17	17	18	18	18	18	19	19	20	21	21	21	21
Pseudomonas putida ATCC 17453	21	8	12	14	15	15	15	16	16	16	16	16	16	16	17	17
Pseudomonas putida Seab04	18	6	12	13	14	14	14	14	14	14	14	14	15	15	16	16
Pseudomonas putida Seab11	19	6	12	14	14	15	16	16	16	16	16	16	16	16	16	16
Pseudomonas sp. Turb25	16	6	9	11	12	12	12	12	12	12	12	12	12	12	12	12
Pseudomonas sp. Turb44	20	4	9	11	13	13	13	13	13	15	15	16	16	16	16	16
Pseudomonas syringae ATCC 19304	32	8	15	20	22	22	22	22	22	22	22	23	23	24	24	24
Pseudomonas syringae ATCC 19310	31	10	14	18	19	19	19	19	19	19	19	20	20	20	20	21
Pseudomonas syringae Seab02	32	10	14	14	16	20	20	20	20	22	22	22	22	22	22	22
Raoultella planticola ATCC 33531	28	11	14	17	17	17	18	18	18	18	19	19	19	19	19	19
Serratia liquefaciens ATCC 12926	28	8	11	12	15	17	17	18	18	18	18	18	18	18	18	18
Serratia marcescens ATCC 274	22	9	12	13	15	16	17	17	17	17	18	18	18	18	18	18
Serratia marcescens Proc7T6	14	4	4	5	8	8	8	8	8	8	8	9	9	9	9	9
Serratia proteamaculans Proc5T6	15	3	3	5	7	9	9	9	9	9	9	9	9	9	9	9
Serratia proteamaculans ProcB1	18	3	8	10	10	12	12	12	12	12	12	12	12	12	12	12
Serratia proteamaculans ProcB4	18	2	6	9	10	11	11	11	11	11	11	11	11	11	11	11
Shewanella algae ATCC 51192	30	9	11	16	17	19	19	19	19	19	19	19	19	19	19	19
Shewanella baltica CECT 323	29	7	11	14	17	20	20	20	20	20	20	20	20	20	20	20
Shewanella putrefaciens ATCC 8071	27	8	12	14	15	18	18	18	18	18	18	18	18	19	19	19
Staphylococcus aureus ATCC 35845	9	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5
Staphylococcus aureus ATCC 9144	10	1	4	4	5	6	6	6	6	6	6	6	6	6	6	6
Staphylococcus aureus Proc1010	13	0	1	2	3	4	4	4	4	4	4	4	4	4	4	4
Staphylococcus epidermidis ATCC 35983	12	3	4	6	7	7	7	7	7	7	7	7	7	7	7	7
Staphylococcus pasteuri 24MF	8	1	3	4	4	4	4	5	5	5	5	5	5	5	5	5
Staphylococcus xylosus ATCC 29971	15	5	7	9	9	9	9	9	9	9	9	9	9	9	9	9
Stenotrophomonas maltophilia 15MF	29	8	10	14	16	16	17	18	18	18	18	18	18	18	18	18
Stenotrophomonas maltophilia 25MC6	29	7	9	10	10	11	11	11	12	13	14	14	14	14	16	16
Stenotrophomonas maltophilia 5PC6	25	6	10	11	11	12	13	13	14	14	14	14	14	14	14	14
Stenotrophomonas maltophilia ATCC 13637	23	8	10	16	18	19	19	19	19	19	19	19	19	19	19	19
Stenotrophomonas maltophilia Seab01	32	9	15	18	20	20	20	20	20	20	20	20	21	22	22	22
Stenotrophomonas maltophilia Seab05	22	8	11	13	14	15	16	16	17	17	17	17	17	17	17	17
Stenotrophomonas maltophilia Seab06	18	7	9	10	12	13	13	13	14	14	14	14	14	14	14	14
Stenotrophomonas maltophilia Seab08	25	13	17	19	19	20	20	20	20	21	21	21	21	21	21	21

Vibrio alginolyticus ATCC 17749	25	5	7	10	13	14	15	16	16	16	16	16	16	16	16	16
Vibrio parahaemolyticus ATCC 17802	27	12	13	15	15	15	16	16	16	16	16	16	16	16	16	17
Vibrio vulnificus ATCC 27562	34	8	13	20	20	21	21	21	21	21	21	22	22	22	23	23