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

DOUGLAS TOMACHEWSKI

UTILIZAÇÃO DE APRENDIZADO DE MÁQUINA PARA CLASSIFICAÇÃO DE BACTÉRIAS ATRAVÉS DE PROTEÍNAS RIBOSSOMAIS

PONTA GROSSA
2017

DOUGLAS TOMACHEWSKI

UTILIZAÇÃO DE APRENDIZADO DE MÁQUINA PARA CLASSIFICAÇÃO DE BACTÉRIAS ATRAVÉS DE PROTEÍNAS RIBOSSOMAIS

Dissertação apresentada ao Programa de Pós-Graduação em Computação Aplicada da Universidade Estadual de Ponta Grossa, como requisito parcial para obtenção do título de Mestre.
Orientação: Prof. Dr. Arion de Campos Júnior
Coorientação: Prof. Dr. Rafael Mazer Etto

PONTA GROSSA
2017

Ficha Catalográfica
Elaborada pelo Setor de Tratamento da Informação BICEN/UEPG

Tomachewski, Douglas
T655 Utilização de aprendizado de máquina para classificação de bactérias através de proteínas ribossomais/ Douglas Tomachewski. Ponta Grossa, 2017.
72f.

Dissertação (Mestrado em Computação Aplicada - Área de Concentração: Computação para Tecnologias em Agricultura), Universidade Estadual de Ponta Grossa.
Orientador: Prof. Dr. Arion de Campos Júnior.
Coorientador: Prof. Dr. Rafael Mazer Etto.

1.Espectrometria de massa. 2.Proteínas ribossomais. 3.Pesos moleculares estimados. 4.Aprendizado de máquina. I.Campos Júnior, Arion de. II. Etto, Rafael Mazer. III. Universidade Estadual de Ponta Grossa. Mestrado em Computação Aplicada. IV. T.

CDD: 006.33

TERMO DE APROVAÇÃO

Douglas Tomachewski

"UTILIZAÇÃO DE APRENDIZADO DE MÁQUINA PARA CLASSIFICAÇÃO DE BACTÉRIAS ATRAVÉS DE PROTEÍNAS RIBOSSOMAIS"

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:

Dr. Rafael Mazer Etto
UEPG

Drª Alaine Margarete Guimarães
UEPG

Dr. Leonardo Magalhães Cruz
UFPR/Curitiba - PR

Ponta Grossa, 04 de setembro de 2017.

Dedico este trabalho aos meus pais Lídia e Celso, e à minha irmã Fabiane, por serem meu alento e fontes de minha dedicação. Também ao meu irmão Maurício (in memorian), que até o fim de sua jornada, nunca deixou de me incentivar.

AGRADECIMENTOS

Agradeço a Deus por guiar-me pelos caminhos da vida, pelos dons a mim concedidos e por todo Seu amor.

Aos meus pais Lídia e Celso e aos meus irmãos Fabiane e Maurício (*in memorian*), por terem me apoiado durante essa e outras jornadas, sempre estando presentes e compartilhando dos meus bons e maus momentos.

Agradeço ao meu orientador Prof. Dr. Arion de Campos Júnior, pela sempre prontidão a me ajudar e repassar o seu conhecimento.

Agradeço ao meu coorientador Prof. Dr. Rafael Mazer Etto, por me conduzir nesta jornada compartilhando seu conhecimento e me mostrando novos horizontes. Saio daqui não só com uma referência de pesquisador, mas também de alguém que ama o que faz.

Agradeço muito a Profª. Drª. Carolina Weigert Galvão por todas as suas orientações e auxílios presentes neste projeto, seu tempo e dedicação foram de fundamental importância para mim.

Agradeço ao Prof. Dr. José Carlos Rocha e a Profª. Drª. Alaine Margarete Guimarães por todas as reuniões e orientações a mim concedidas, seus conhecimentos e prestimosidades formaram um conhecimento-chave para realizar este trabalho.

Agradeço à Universidade Estadual de Ponta Grossa e ao programa de Pós-Graduação em Computação Aplicada por me aceitarem como discente e me oferecerem todo amparo e infraestrutura.

Aos professores do programa de Pós-Graduação em Computação Aplicada por compartilhar seus conhecimentos e dedicarem-se à docência deste programa.

Agradeço ao Laboratório de Biologia Molecular Microbiana (LABMOM) por me acolher e confiar a mim seus dados, para que eu pudesse realizar meus estudos.

Agradeço à Universidade Federal do Paraná, em especial a Profª. Drª. Maria Berenice Steffens e ao Prof. Dr. Leonardo Cruz, por me receber em seu programa de Pós-Graduação em Bioinformática oferecendo-me espaço para adquirir novos conhecimentos e trocar experiências.

Aos meus amigos do Lab14, Fábio, Giancarlo, Rodrigo, Francisco e Renann por vivenciarem comigo esta jornada e dividirem suas experiências.

Aos meus colegas do Mestrado e do INFOAGRO, agradeço pela amizade e incentivo.

Muito obrigado!

“É muito melhor lançar-se em busca de conquistas grandiosas, mesmo expondo-se ao fracasso, do que alinhar-se com os pobres de espírito, que nem gozam muito nem sofrem muito, porque vivem numa penumbra cinzenta, onde não conhecem nem vitória, nem derrota.” (Theodore Roosevelt)

RESUMO

A identificação de microrganismos, nas áreas da saúde e agricultura, é essencial para compreender a composição e o desenvolvimento do meio. Novas técnicas estão buscando identificar estes microrganismos com mais acurácia, rapidez e com menor custo. Uma técnica cada vez mais estudada e utilizada atualmente é a identificação de microrganismos através de espectros de massa, gerados por uma espectrometria de massa. Os espectros de massa são capazes de gerar um perfil para reconhecimento de um microrganismo, utilizando os picos referentes às mais abundantes massas moleculares registradas nos espectros. Analisando os picos pode-se designar um padrão, como uma impressão digital, para reconhecer um microrganismo, esta técnica é conhecida como PMF, do inglês *Peptide Mass Fingerprint*. Outra forma de identificar um espectro de massa, é através dos picos que são esperados que se apresentem no espectro, modelo qual este trabalho utilizou. Para prever os picos esperados no espectro, foram calculados os pesos moleculares estimados de proteínas ribossomais. Essas proteínas são denominadas *house keeping*, ou seja são presentes para o próprio funcionamento celular. Além de apresentarem grande abundância no conteúdo procariótico, elas são altamente conservadas, não alterando sua fisiologia para diferentes meios ou estágios celulares. Os pesos estimados formaram uma base de dados presumida, contendo todas as informações obtidas do repositório do NCBI. Esta base de dados presumida foi generalizada para taxonomia a nível de espécie, e posteriormente submetida à um aprendizado de máquina. Com isso foi possível obter um modelo classificatório de microrganismos baseado em valores de proteínas ribossomais. Utilizando o modelo gerado pelo aprendizado de máquina, foi desenvolvido um *software* chamado Ribopeaks, capaz classificar os microrganismos a nível de espécie com acurácia de 94.83%, considerando as espécies correlatas. Também foram observados os resultados a nível taxonômico de gênero, que obteve 98.69% de assertividade. Valores de massas moleculares ribossomais biológicas retiradas da literatura também foram testadas no modelo obtido, obtendo uma assertividade total de 84,48% para acertos em nível de espécie, e 90,51% de acerto em nível de gênero.

Palavras-chave: Espectrometria de Massa, Proteínas Ribossomais, Pesos moleculares estimados, Aprendizado de Máquina

ABSTRACT

Identification of microorganisms in health and agriculture areas is essential to understand the composition and development of the environment. New techniques are seeking to identify these microorganisms with more accuracy, speed and at a lower cost. Nowadays, a technique that is increasingly studied and used is the identification of microorganisms through mass spectra, generated by mass spectrometry. The mass spectra are able to generate a recognition profile from a microorganism, using the referring peaks to the most abundant molecular masses recorded in the spectrum. By analyzing the peaks, it is possible to designate a pattern, such as a fingerprint, to recognize a microorganism; this technique is known as the Peptide Mass Fingerprint (PMF). Another way to identify a mass spectrum is through the peaks that are expected to appear in the spectrum, which model this work used. To predict the expected peaks in the spectrum, the estimated molecular weights of ribosomal proteins were calculated. These proteins are responsible for the cellular functioning itself, so-called housekeeping. Besides they being abundant in the prokaryotic content, they are highly conserved, not altering their physiology to different environments or cell stage. The estimated weights formed a presumed database, containing all the information obtained from the NCBI's repository. This presumed database was generalized at the specie level and later submitted to a machine learning algorithm. With this, it was possible to obtain a microorganism's classificatory model based on ribosomal proteins values. Using the generated model by the machine learning, a software called Ribopeaks was developed to classify the microorganisms at the specie level with an accuracy of 94.83%, considering the related species. It was also observed the results at genus level, which obtained 98.69% of assertiveness. Values of biological ribosomal molecular masses from the literature were also tested in the acquihired model, obtaining a total assertiveness of 84.48% at the specie level, and 90.51% at the genus level.

Keywords: Mass Spectrometry, Ribosomal Proteins, Estimated Molecular Weights, Machine Learning

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

30S	Subunidade menor de um ribossomo em procariotos
50S	Subunidade maior de um ribossomo em procariotos
ACO	*Ant Colony Optimization*
API	*Application Programming Interface*
BPCV	Bactérias Promotoras do Crescimento Vegetal
Da	Daltons, unidade de medida de m/z
DNA	Ácido Desoxirribonucleico
EM	Espectrometria de Massa
H	Hidrogênio
IA	Inteligência Artificial
kDa	Unidade correspondente a mil Daltons
Kernels	Núcleos; Médias
LABMOM	Laboratório de Biologia Molecular Microbiana da UEPG
LC-MS/MS	Uma combinação de Cromatografia Líquida (LC) e Espectrometria de Massa (MS)
m/z	Relação massa/carga da amostra ionizada, disposta no eixo x do espectro
MALDI-TOF	*Matrix-assisted laser desorption/ionization time-of-flight*
MLST	Análise de sequências multi-locais
mRNA	Ácido ribonucleico mensageiro
NCBI	Centro Nacional de Informação Biotecnológica
OH	Hidroxila
PMF	*Peptide mass fingerprint*
PM	Peso Molecular
RNA	Ácido Ribonucleico
ROC	*Receiver Operating Characteristic*
rRNA	Ácido Ribonucleico ribossomal
SVM	*Support Vector Machine*

SUMÁRIO

1 INTRODUÇÃO

Novas metodologias para realizar a identificação de microrganismos que sejam mais rápidas, baratas e confiáveis são desafios constantes da pesquisa científica. A identificação bacteriana é essencial para diversas áreas, como as áreas da saúde e da agricultura.

O desenvolvimento de metodologias mais rápidas e eficientes na identificação de bactérias, apesar de prioritariamente visar a resolução de problemas clínicos, pode contribuir também para a identificação de novos microrganismos, como as Bactérias Promotoras do Crescimento Vegetal (BPCV). Isso vai ao encontro das demandas da agricultura moderna, uma vez que um dos grandes desafios do século XXI será alimentar aproximadamente 10 bilhões de pessoas em 2050 e ao mesmo tempo reduzir o impacto ambiental e os custos da produção agrícola (FOLEY et al., 2011). Nesse contexto, as BPCV ganham destaque, pois são responsáveis por fornecer nutrientes e fitormônios essenciais para o desenvolvimento da planta, substituindo a necessidade do uso de agroquímicos (VESSEY, 2003). Além da sua importância na produtividade agrícola, é sabido que essas bactérias ambientais são essenciais para a manutenção dos ciclos biogeoquímicos. A necessidade de desenvolver metodologias mais rápidas e menos custosas para identificar novas BPCV é decorrente da diversidade dessas bactérias no interior da planta ou em sua rizosfera (BULGARELLI et al., 2013). Dados estatísticos indicam que 1g de solo pode conter de 2×10^3 a $8{,}3 \times 10^6$ espécies bacterianas diferentes, onde apenas 1 a 5% são conhecidas (GANS et al., 2005 e SCHLOSS et al., 2006).

Uma tecnologia que está ganhando cada vez mais espaço nos laboratórios e nas pesquisas, é a Espectrometria de Massa (EM). A Espectrometria de Massa do tipo MALDI-TOF (matriz assistida por dessorção e ionização por tempo de voo, do inglês *Matrix-assisted Dessortion/Ionization time-of-flight*), que gera espectros de massas das amostras analisadas e é amplamente utilizada. Isto ocorre devido a fatores como a facilidade na preparação da amostra, que dispensa muito trabalho laboratorial, pois não é necessário extrair o material genético do microrganismo para obter um perfil de dados necessário para sua identificação. Outro fator positivo é o tempo reduzido de obtenção dos resultados, em relação a métodos de cultivo ou moleculares tradicionais. Além disso, o custo envolvido no sequenciamento do material genético extraído de um microrganismo, é maior do que o custo de preparo de uma amostra para a EM (GARCÍA et al., 2012).

Os espectros de massa gerados pelo MALDI-TOF podem ser analisados de duas maneiras. A primeira compara os espectros entre si, buscando seus picos em comum e semelhança entre eles. Na literatura essa metodologia é conhecida como PMF, do inglês *Peptide Mass Fingerprint* (Impressão Digital de Massa Peptídica). Outra maneira utilizada é calculando e prevendo pesos moleculares presentes nos espectros e procurando por biomarcadores capazes de designar um padrão para identificar um microrganismo (TAMURA; HOTTA e SATO, 2013).

No entanto, estipular biomarcadores não é uma tarefa trivial, alguns polipeptídios não são expressos constitutivamente e portanto, a sua presença no perfil do espectro de massa pode variar dependendo da condição fisiológica da bactéria. Teramoto et al. (2007) abordam esta questão e apresenta o uso das proteínas ribossomais como biomarcadoras em espectros de MALDI-MS. A justificativa para o uso de tais proteínas, baseia-se na alta conservação que as proteínas ribossomais apresentam e na sua expressão constitutiva em diferentes condições fisiológicas, independente de meios de crescimento ou ciclo celular.

O aprendizado de máquina, vêm ganhando destaque em análises de dados de EM. Trabalhos como de De Bruyne et al. (2010), Villanueva et al. (2004), entre outros, mostram como a predição de um modelo de aprendizado de máquina, pode aumentar a capacidade classificatória de diferentes amostras analisadas na EM.

Neste trabalho foi utilizado a predição de biomarcadores em espectros, procurando especificamente por picos associados a proteínas ribossomais. Utilizou-se posteriormente o aprendizado de máquina para reconhecer padrões, e obter um modelo de reconhecimento bacteriano baseado nos biomarcadores ribossomais.

Em sequência são apresentados os Objetivos deste trabalho na seção 2, seguidos da Revisão de Literatura na seção 3. Serão abordados na Revisão de Literatura trabalhos relacionados a EM e seu funcionamento, bem como as formas de classificação de espectros. Mais informações e demais trabalhos sobre proteínas ribossomais também são discorridos na Revisão de Literatura, seguidos pelas definições e trabalhos relacionados ao aprendizado de máquina. Na seção 4 Materiais e Métodos são apresentados a aquisição dos dados, a criação da base de dados presumida, a construção do modelo, o software de classificação e os perfis de espectro de gênero. Na seção 5 Resultados e Discussões são discutidos as abordagens utilizadas e apresentado os resultados obtidos. Por fim, na seção 6 são expostas as Conclusões deste trabalho.

2 OBJETIVOS

2.1 OBJETIVO GERAL

Gerar um classificador taxonômico de bactérias baseado em espectros de massa ribossomais virtuais.

2.2 OBJETIVOS ESPECÍFICOS

- Construir uma calculadora para determinar os pesos moleculares a partir das sequências de aminoácidos das proteínas ribossomais, considerando as modificações pós-traducionais;
- Criar uma base de dados com os pesos moleculares presumidos das proteínas ribossomais;
- Criar um software para classificação baseado no modelo gerado pelo aprendizado de máquina.

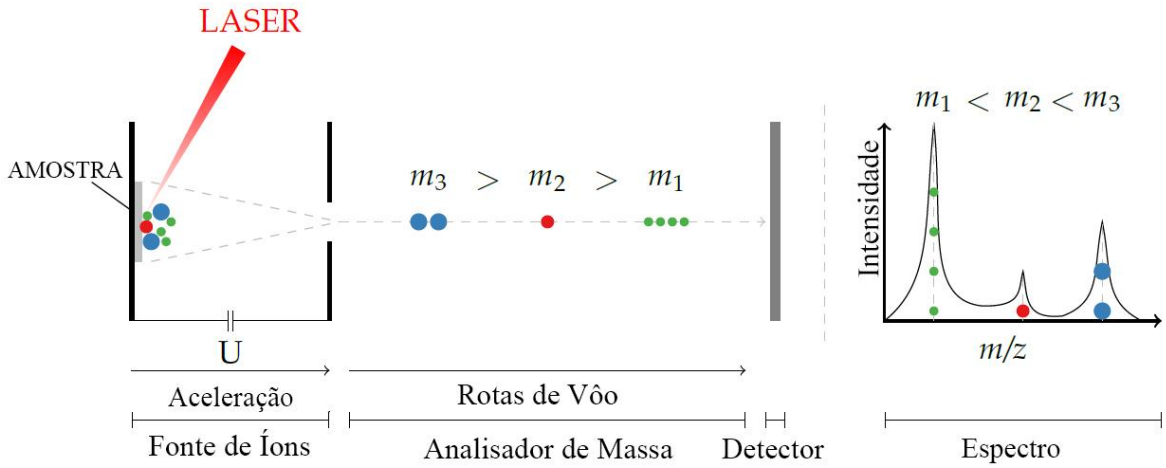
3 REVISÃO DE LITERATURA

3.1 ESPECTROMETRIA DE MASSA DO TIPO MALDI-TOF MS

A Espectrometria de Massa é uma técnica que consiste em caracterizar as moléculas pela medida da relação massa/carga (m/z) de seus íons (WILLARD et al., 1988). Nos últimos anos, a EM do tipo MALDI-TOF vem sendo utilizada para identificação de bactérias de importância clínica de modo rápido e preciso (GARCÍA et al., 2012).

Na EM de células bacterianas inteiras (GIBB e STRIMMER, 2012), o espectro de massa gerado é o resultado do perfil de expressão proteica da célula em uma determinada condição fisiológica (Figura 1).

FIGURA 1 – Representação do funcionamento do Espectrômetro de Massa do tipo MALDI-TOF



Legenda: Esquema representando o funcionamento do MALDI-TOF até a obtenção de um espectro de massa, a ionização das amostras, o voo das diferentes macromoléculas ionizadas pelo tubo de voo (*m1, m2 e m3*), e a geração dos picos no espectro quando as macromoléculas atingem o detector.
Fonte: Gibb (2014).

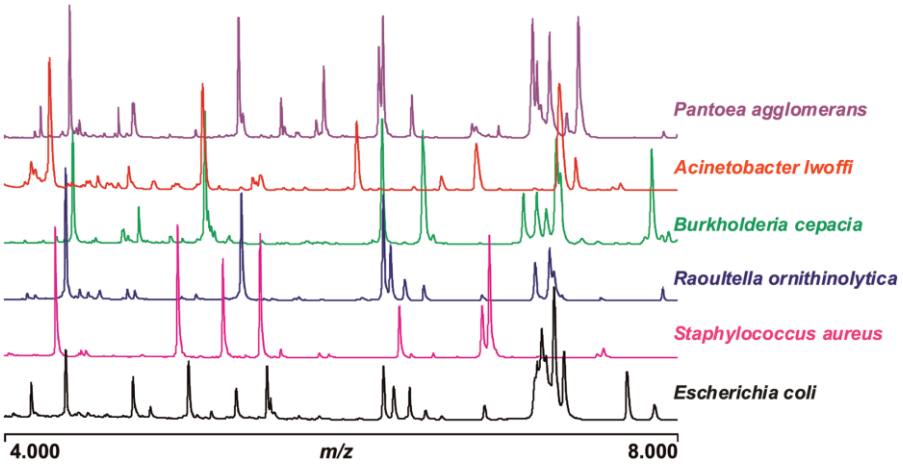
Na Figura 1 é mostrado como o espectro é obtido por MALDI-TOF. As amostras depositadas em uma placa recebem um *laser* e são ionizadas em uma câmara dentro do aparelho. Em seguida, o *laser* disparado sobre elas conduz as moléculas ionizadas através de um tubo mantido em alto vácuo. Estas moléculas voam pelo tubo até atingirem um detector,

que capta os sinais e gera o espectro. Moléculas menores chegam primeiro ao detector, enquanto as maiores têm seu tempo de voo mais prolongado. A intensidade do pico é indicada no eixo y e o eixo do espectro indica o tempo de voo, que é diretamente proporcional a razão *m/z*.

Segundo García et al. (2012), a identificação bacteriana a partir de EM é uma metodologia promissora e será cada vez mais utilizada nos diferentes setores da sociedade. A identificação bacteriana por EM do tipo MALDI-TOF comparada a outros métodos tradicionais, como sequenciamento de DNA ou testes químicos, é muito vantajosa pela economia de dinheiro e tempo. O uso das análises por EM são cinco vezes menos custosa e demora em média 6 minutos. O tempo gasto para identificação de bactérias por métodos tradicionais pode demorar de horas até dias para ser finalizada. Portanto, o alto rendimento e a mínima preparação das amostras, somada à velocidade da aquisição dos dados e da automação no processo de análise, faz da EM do tipo MALDI-TOF uma poderosa metodologia para a identificação de bactérias (STETS et al., 2013).

Dois métodos para identificação bacteriana são utilizados em dados de EM. O primeiro método consiste em comparar diferentes espectros de massa, permitindo a identificação de picos comuns entre os espectros. Isto torna possível o agrupamento das bactérias analisadas e posteriormente, a classificação taxonômica. Esses picos comuns são conhecidos como PMF, do inglês *peptide mass fingerprint*, que servem como impressões digitais da bactéria (Figura 2). Este primeiro método identifica bactérias até o nível de espécie, independente do meio ou habilidades do analista, pois é composto somente dos registros das cargas massas ionizadas (TAMURA; HOTTA e SATO, 2013).

FIGURA 2 – Exemplos de espectros



Legenda: Seis diferentes exemplos de espectros de massas, mostrando os diferentes picos de carga massa entre as espécies.
Fonte: García et al. (2012).

Muitos trabalhos já usaram o método de comparação de espectros de massa para classificação bacteriana, como foram os casos dos softwares SPECLUST (JOHANSSON et al., 2007), MALDIquant (GIBB e STRIMMER, 2012), MS Analyser (SANTOS et al., 2013) e MS-ML Studio (BRIM et al., 2015). Estes trabalhos abordam etapas como o pré-processamento, seleção de picos e agrupamento de espectros, posteriormente analisando a identificação das bactérias. No trabalho realizado por Müller et al. (2002), foi utilizado um *scanner* molecular proposto por Binz et al. (1999). Nessa metodologia Müller et al. (2002), fizeram uma análise combinada entre um gel de eletroforese e as impressões digitais dos peptídeos (PMF), para realizar o agrupamento dos espectros.

Schmidt et al. (2003) também utilizaram PMF's para identificar picos de massa de peptídeos. Após realizar eletroforese bidimensional e comparar os picos com registros em um banco de dados, perceberam que uma drástica quantidade de picos não era assimilada. Estes picos não assimilados foram identificados no trabalho como picos provenientes de contaminantes, *spots* vizinhos ou mesmo por modificação pós-traducional.

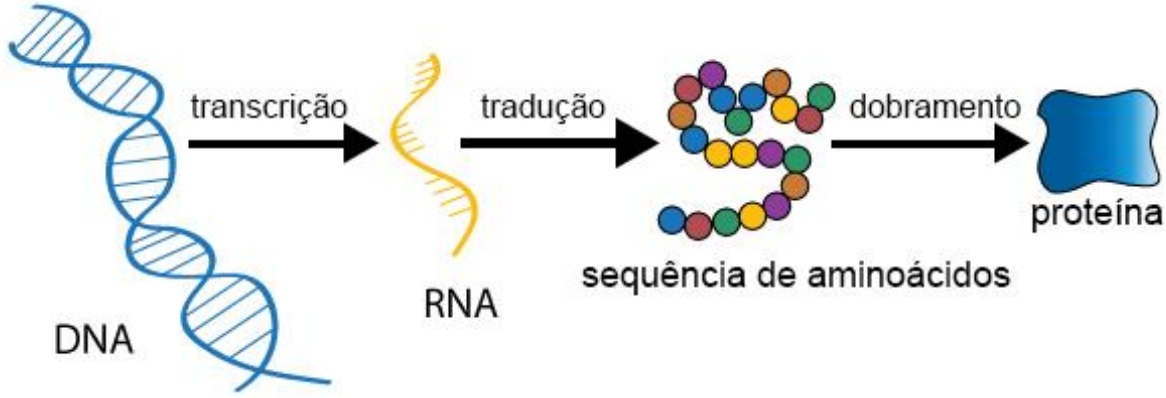
Tibshirani et al. (2004) propuseram a utilização da espectrometria de massa para identificar mais rapidamente o câncer de ovário. A avaliação proposta envolve comparar as amostras de espectros de pacientes saudáveis com aqueles com a doença, mostrando que a probabilidade da similaridade e dissimilaridade entre eles determinava a diferença entre os dois grupos.

Beer et al. (2004) agruparam espectros gerados em LC-MS/MS (*Liquid chromatography–mass spectrometry*) para reduzir a quantidade do volume de dados gerados a partir dessas análises. Além de agrupar os espectros para economia de espaço, também foi possível identificar mais peptídeos quando realizado a melhoria dos espectros.

Monigatti e Berndt (2005) sugerem que a partir de grandes bases de dados de espectros, tenham-se espectros consensuais. Estes espectros tenderiam a diminuir os falsos positivos em análises, e ajudar a entender posteriormente o comportamento de peptídeos na EM.

O segundo método utilizado para identificação bacteriana consiste em encontrar os biomarcadores representados pelos picos de massas moleculares presentes nos espectros. Neste método procura-se calcular e predizer a massa molecular ionizada (*m/z*), baseado nas sequências de aminoácidos traduzidas pela informação de genes provenientes de genomas sequenciados dos microrganismos (DEMIREV et al., 1999). O processo de como um gene transforma-se em uma proteína, é demonstrado na Figura 3.

FIGURA 3 – Processo de expressão de uma proteína



Legenda: Processo de expressão de uma proteína, começando na transcrição do gene para um mRNA, a tradução do RNA mensageiro para formação da cadeia polipeptídica, e o natural dobramento da cadeia formando a proteína final.
Fonte: **Biosocial Methods Collaborative** – Universidade de Michigan. Disponível em: <http://biosocialmethods.isr.umich.edu/epigenetics-tutorial/epigenetics-tutorial-gene-expression-from-dna-to-protein>. Acesso em: 26 jul. 2017.

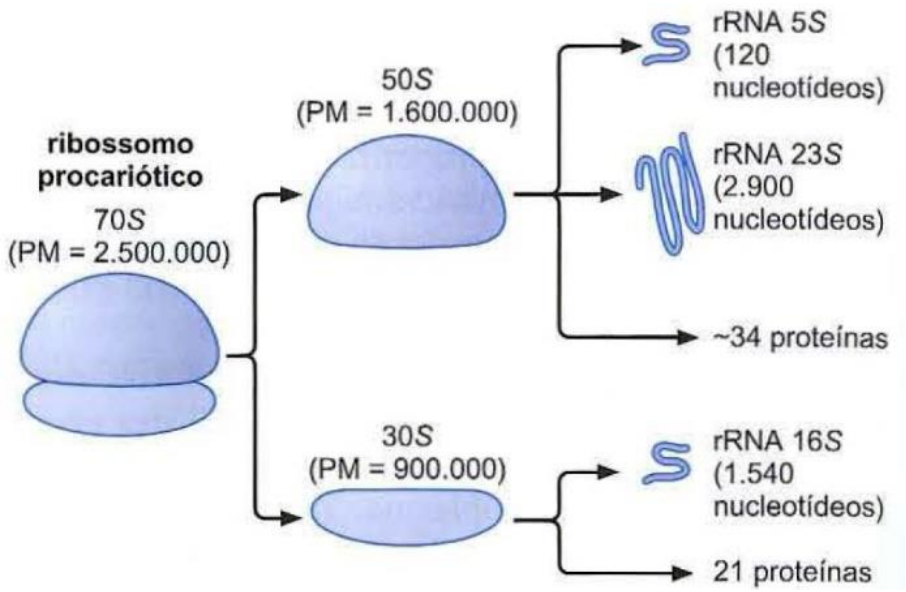
Este método tem maior potencial de identificação bacteriana, uma vez que picos específicos da estirpe ou espécie já foram preditos. Alguns pontos importantes sobre esta metodologia devem ser considerados, como haver o genoma ou a sequência do gene já sequenciado, obter uma informação genômica acurada, e as modificações pós-traducionais que ocorrem nas proteínas (TAMURA; HOTTA e SATO, 2013).

3.2 PROTEÍNAS RIBOSSOMAIS

O ribossomo é o elemento molecular responsável por realizar a síntese proteica, e são encontrados em organismos procarióticos e eucarióticos. O ribossomo procariótico (70S) é

formado por duas subunidades, elas são conhecidas como subunidade maior (50S) e subunidade menor (30S). Em sua composição cada subunidade é constituída por RNA ribossomal, ou rRNA, e por proteínas ribossomais (Figura 4) (WATSON et al., 2015).

FIGURA 4 – Composição do ribossomo procariótico



Legenda: A composição proteica e os rRNAs das diversas subunidades são indicadas. Também são apresentados os tamanhos dos rRNAs, o peso molecular, ou massa molecular, (PM) e o número de proteínas ribossomais de cada subunidade.
Fonte: Watson et al. (2015).

As proteínas ribossomais são um complexo de proteínas conhecidas como *housekeeping*, ou seja, proteínas que desempenham funções vitais da própria célula. Elas encontram-se em grande quantidade, e estão presentes nas subunidades maiores e menores que formam os ribossomos. Estas proteínas são altamente conservadas, ou seja, não sofrem grande alteração em sua sequência de aminoácidos (TERAMOTO et al., 2007). Assim, as proteínas ribosomais podem ser consideradas biomarcadores confiáveis para identificação bacteriana.

No trabalho apresentado por Tamura, Hotta e Sato (2013), foi utilizada a técnica de predição de pesos moleculares para as proteínas ribossomais do operon *S10-spc-alpha*, que compreendem metade das proteínas ribossomais e são altamente conservadas. O objetivo do trabalho foi discriminar espécies e estirpes do gênero *Pseudomonas,* posteriormente também aplicado com sucesso para diferenciar a estirpe patógena de *Pseudomonas syringae*, comumente associada a doenças de plantas. Também foi demonstrado a eficiência da técnica,

caracterizando o grupo de *Lactobacillus casei* e discriminando bactérias dos gêneros *Bacillus* e *Sphingopyxis*.

Demirev et al. (1999) também demonstraram o uso de biomarcadores para correta identificação da *Bacillus subtilis* e da *Pseudomonas syringae*. Os testes foram realizados em espectros de massa com mistura de microrganismos, espectros de organismos em diferentes estágios de crescimento, e em espectros provenientes de outros laboratórios. Os autores concluíram que a técnica foi eficiente para identificação de bactérias.

Teramoto et al. (2007) aplicaram a mesma técnica para identificar estirpes da *Pseudomonas putida*, definindo 43 proteínas ribossomais biomarcadoras para realizar a identificação. O microrganismo escolhido, *Pseudomonas putida*, é conhecido por ter uma classificação muito mais complicada que as demais estirpes de outras bactérias. Além disso, os autores relatam que sua eficiência é comparável à de outra metodologia robusta baseada na subunidade do gene B da DNA girase, sugerindo que a metodologia pode ser aplicada para novos isolados de bactérias.

Ziegler et al. (2015) utilizaram proteínas ribossomais biomarcadoras para caracterizar estirpes mais comuns de bactérias que são promotoras do crescimento vegetal. Estes microrganismos são conhecidos por nodular raízes de plantas, gerando nódulos que fixam o nitrogênio da atmosfera. Esta relação simbiótica promove maior desenvolvimento do vegetal, aumentando a produtividade e até mesmo substituindo alguns fertilizantes. Dispensando de altos custos laboratoriais e tempos de preparação amostral, a técnica por espectrometria de massa mostrou-se eficaz em 116 estirpes de gêneros rizobiais. Neste estudo foram reportadas 13 proteínas ribossomais biomarcadoras, suficientes para diferenciar os microrganismos envolvidos.

Estes trabalhos mostraram que proteínas ribossomais podem ser biomarcadores eficientes para a identificação bacteriana, uma vez que apresentam alto nível de conservação, reduzido número de aminoácidos e expressão constitutiva. Outras informações ribossomais atualmente já foram utilizadas para identificar bactérias, como é o caso do 16S rRNA ribossomal. Este trecho de código genético presente na subunidade menor do ribossomo, também apresenta informações conservadas. A determinação da sequência de nucleotídeos do gene 16S rRNA é amplamente utilizada para identificação de microrganismos a nível de gênero. Esse método apresenta alto custo e requer uma trabalhosa preparação amostral (GARCÍA et al., 2012).

3.3 APRENDIZADO DE MÁQUINA COM DADOS DE MALDI-TOF MS

O aprendizado de máquina emergiu nas últimas décadas como um recurso mais sofisticado para solução de problemas a serem tratados, funcionando de forma mais autônoma a outros recursos programados com base em conhecimento de especialistas. Um algoritmo de aprendizado de máquina deve oferecer uma solução através de uma hipótese ou função, baseado em uma experiência previamente fornecida. Espera-se também que um algoritmo de aprendizado de máquina possa analisar informações, de forma a otimizar o seu modelo de resultado, assim como reconhecer e organizar um conhecimento novo (FACELI et al., 2011).

Mitchell (1997) define aprendizado de máquina como: *"A capacidade de melhorar o desempenho na realização de alguma tarefa por meio da experiência."*

Dentre os algoritmos presentes no aprendizado de máquina, define-se os algoritmos em supervisionados e os não supervisionados. A diferença entre estes está na presença de um atributo meta na base de dados fornecida como conhecimento ao algoritmo. No aprendizado de máquina supervisionado tem-se as tarefas de classificação e regressão, para o aprendizado de máquina não supervisionado tem-se as tarefas de agrupamento, associação e sumarização.

No contexto deste trabalho, onde tem-se na maioria das vezes a taxonomia de um microrganismo como atributo meta, é mais comum encontrar na literatura o uso de algoritmos de aprendizado de máquina supervisionados que realizam a tarefa de classificação dos objetos. Os algoritmos mais comuns para a tarefa de classificação são: 1) algoritmos baseados em distâncias, como o vizinho mais próximo e *k*-NN. 2) algoritmos baseados em casos. 3) algoritmos probabilísticos, como a rede bayesiana e o *Naïve Bayes*. 3) algoritmos baseados em procura, como árvores de decisão e regras de decisão. 4) algoritmos baseados em otimização, como as redes neurais artificiais e SVM (*Support Vector Machine*). Além de ainda haver a utilização de modelos múltiplos preditivos, que realizam a combinação de dois ou mais algoritmos de aprendizado (FACELI et al., 2011).

Há na literatura trabalhos que abordam o aprendizado de máquina e dados de espectros gerados por MALDI-TOF, os quais serão descritos a seguir, porém nenhum deles usa especificamente proteínas ribossomais para identificação bacteriana.

No trabalho de De Bruyne et al. (2010) foi apresentado uma metodologia desde a preparação das amostras até a obtenção dos espectros. Estes espectros foram pré-processados e extraídos para identificação das espécies dos gêneros *Leuconostoc*, *Fructobacillus* e

Lactococcus. Utilizando as bibliotecas padrões de identificação do MALDI-TOF, os autores obtiveram um acerto de 84% na identificação das espécies de *Leuconostoc* e *Fructobacillus*, e de 94% para espécies do gênero *Lactococcus*. No entanto, após utilizarem duas técnicas de aprendizado de máquina, SVM e *Random Forest*, a assertividade na classificação dos dados foi de 94% e 98% na identificação de espécies de *Leuconostoc* e *Fructobacillus*, respectivamente.

Villanueva et al. (2004) utilizaram dados de espectros de massa, extraídos de alcances de 0.8 a 15 kDa, para identificar tumores cerebrais através de polipeptídios presentes em amostras de soro coletados dos pacientes. Aproximadamente 400 pesos moleculares foram reunidos dos espectros, dos quais 274 mostraram-se suficientes para distinção da doença. Estas informações foram submetidas ao aprendizado de máquina utilizando o algoritmo k-NN, o qual gerou um modelo capaz de distinguir com precisão de 96,4% as amostras entre normal e doente.

Para o grande número de picos encontrado em Villanueva et al. (2004), o trabalho de Ressom et al. (2007) sugere o uso de outro algoritmo de aprendizado de máquina para realizar a classificação das amostras. O algoritmo indicado foi o SVM para obter o modelo de classificação, com este foi obtido 94% de sensibilidade e 100% de especificidade na análise das amostras através do modelo. Outras técnicas no pré-processamento dos espectros também foram abordadas, um exemplo foi o próprio pré-processamento dos espectros não tratados obtidos do MALDI-TOF, sendo realizado a redução de ruídos, suavização, correção de eixo e normalização dos espectros. Além disto foi proposto pelos autores uma redução na dimensionalidade dos biomarcadores propostos para apenas 8 destes, para isso foi utilizado um algoritmo bioinspirado em colônia de formigas, o ACO (*Ant Colony Optimization*), para a seleção dos biomarcadores.

No trabalho de Hettick et al. (2006), estirpes de *Mycobacteria tuberculosis*, *M. avium*, *M. intracellulare* e *M. kansaii* foram classificadas utilizando espectros de MALDI e os algoritmos de análise de discriminação linear e *Random Forest*. O melhor resultado apresentado foi com o algoritmo *Random Forest*, apresentando uma taxa de erro de 0,023 e utilizando 18 pesos moleculares para realizar a classificação.

Outra abordagem utilizada por Ulintz et al. (2006) foi o uso dos algoritmos *Boosting* e *Random Forest* para classificar espectros de bactérias obtidas por MALDI-TOF, diferindo verdadeiros positivos de falsos positivos, determinados por outros algoritmos. Foi demonstrado a eficiência dos algoritmos em bases públicas, por meio de gráfico da curva ROC (*Receiver Operating Characteristic*). Outros algoritmos comparados foram o PeptideProphet e o SVM.

Por fim, baseado na revisão de literatura apresentada este trabalho utilizou a técnica de predição de massas moleculares, tendo como biomarcadores definidos as proteínas ribossomais. Para realizar o aprendizado de máquina, foi utilizado o algoritmo probabilístico *Naïve Bayes* com opção *kernels estimator* (JOHN e LANGLEY, 1995).

4 MATERIAIS E MÉTODOS

Para construir um modelo baseado em aprendizado de máquina, inicialmente foi definida a dimensionalidade da estrutura do conhecimento a ser desenvolvido. Para isso, utilizou-se o conceito apresentado em Tamura; Hotta e Sato (2013), no qual a abordagem baseia-se no *S10-spc-alpha operon* de microrganismos procarióticos para caracterização da *Pseudomonas syringae*.

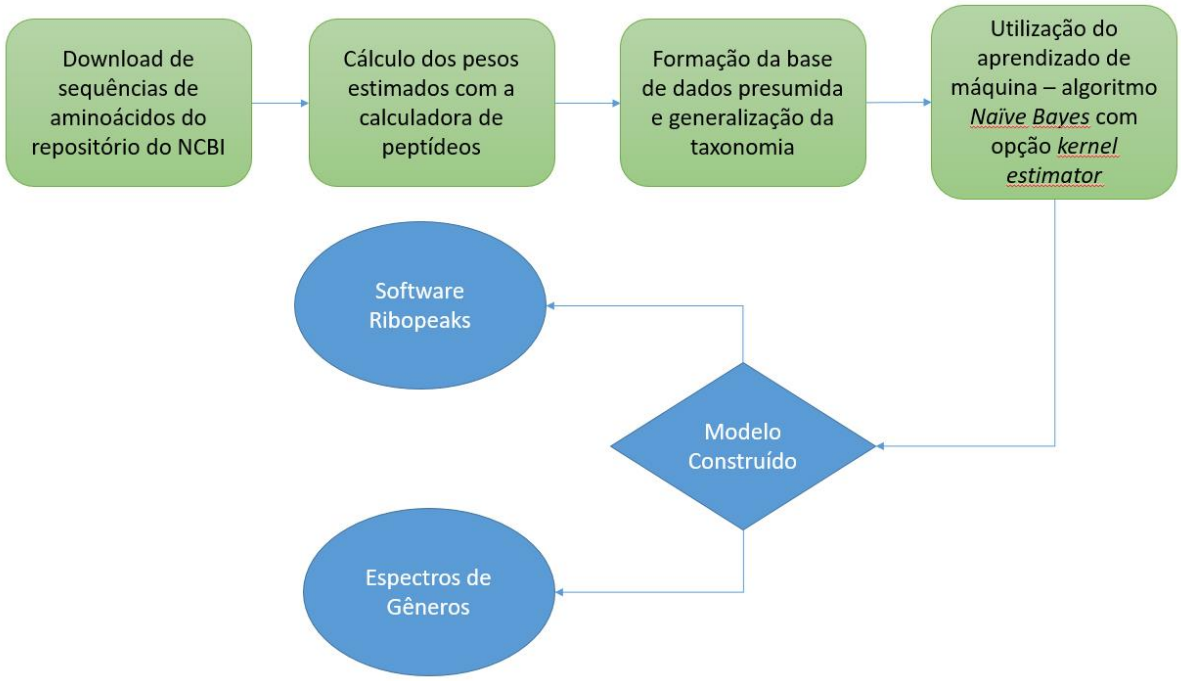
Considerando-se o alto grau de conservação das macromoléculas ribossomais entre as bactérias, foram avaliadas a efetividade das proteínas ribossomais para a identificação bacteriana. Assim, foi necessário construir uma base de dados com todas as proteínas ribossomais comumente presentes nas subunidades maior e menor do ribossomo, respectivamente conhecidas como 50S e 30S.

As 60 proteínas ribossomais selecionadas foram: L1, L2, L3, L4, L5, L6, L7, L7a, L7ae, L7/L12, L9, L10, L11, L12, L13, L14, L15, L16, L17, L18, L19, L20, L21, L22, L23, L24, L25, L27, L28, L29, L30, L31, L32, L33, L34, L35, L36, S1, S2, S3, S4, S5, S6, S7, S8, S9, S10, S11, S12, S13, S14, S15, S16, S17, S18, S19, S20, S21, S22 e S31e. A proteína L8 não está presente, por ser um pentâmero formado pelas proteínas L7 e L10. A proteína ribossomal L26 também não se faz presente, por apresentar a sequência idêntica de aminoácidos da S20. E proteínas ribossomais da S23 a S30, são exclusivas de organismos eucarióticos (STELZL et al., 2001).

O fluxograma mostrado na Figura 5 demonstra os principais passos realizados neste trabalho para obtenção do modelo gerado pelo aprendizado de máquina.

Nos próximos tópicos serão mostrados cada um desses passos, desde a aquisição dos dados no repositório do NCBI, a criação da calculadora de peptídeos considerando as modificações pós-traducionais, a criação da base de dados presumida utilizando as proteínas ribossomais, o uso do aprendizado de máquina para criação de um modelo de classificação, o *software* desenvolvido utilizando o modelo obtido para realizar a classificação de microrganismos, até o uso do mesmo modelo obtido do aprendizado de máquina para geração de espectros visuais que representam gêneros ou espécies.

FIGURA 5 – Fluxograma da metodologia



Legenda: Principais passos realizados para obtenção do modelo do aprendizado de máquina.
Fonte: O autor.

4.1 AQUISIÇÃO DOS DADOS

Todos os dados referentes a proteínas ribossomais (30S e 50S) foram obtidas do NCBI (Centro Nacional de Informação Biotecnológica), presentes no banco de dados de proteínas, na data de 13/06/2016 através da API (Interface de aplicação a programação, do inglês *Application Programming Interface*) *E-utilities*, disponibilizada pelo próprio repositório. Foram obtidas 2.807.341 sequências aminoácidos, utilizando o formato de arquivo ".fasta" para *download*.

Os dados adquiridos foram armazenados em banco de dados Mysql para posterior processamento das informações. A Tabela 1 mostra a estrutura das informações utilizada para armazenar os dados obtidos, sendo "*cod*" a numeração padrão do registro no banco de dados, "*ncbi_id*" número de identificação do registro no repositório, "*source*" referência de origem da sequência, "*source_id*" número de identificação da referência de origem da sequência, "*description*" informação da proteína, "*taxonomy*" taxonomia pertencente à bactéria, "*likely_genus*" gênero relativo ao registro, "*seq_aa*" sequência de aminoácidos.

TABELA 1 – Estrutura de dados utilizada para armazenar dados obtidos do NCBI

Nome do campo	**Tipo de dado**
cod	int(11)
ncbi_id	varchar(21)
source	varchar(17)
source_id	varchar(21)
description	varchar(177)
taxonomy	varchar(177)
likely_genus	varchar(177)
seq_aa	text

Fonte: O autor.

4.2 CALCULADORA DE PEPTÍDEOS

Os dados armazenados na Tabela 1 foram convertidos em pesos moleculares estimados, utilizando uma calculadora de peptídeos, desenvolvida para esse projeto. A calculadora de peptídeos analisa as sequências de aminoácidos fornecidas, e através de pesos isotópicos médios estimados (Seção 5.4) de cada aminoácido (Tabela 2), calcula o peso molecular total da proteína informada.

TABELA 2 – Pesos de aminoácidos utilizados para cálculo do peso molecular das proteínas

Aminoácido	**Sigla**	**Letra na sequência**	**Peso mol. médio (Da)**
Alanina	Ala	a	71,0788
Arginina	Arg	r	156,1876
Asparagina	Asn	n	114,1039
Ácido aspártico	Asp	d	115,0886
Cisteína	Cis	c	103,1448
Ácido glutâmico	Glu	e	129,1155
Glutamina	Gln	q	128,1308
Glicina	Gli	g	57,0520
Histidina	His	h	137,1412
Isoleucina	Ile	i	113,1595
Leucina	Leu	l	113,1595
Lisina	Lis	k	128,1742
Metionina	Met	m	131,1963
Fenilalanina	Fen	f	147,1766
Prolina	Pro	p	97,1167
Serina	Ser	s	87,0782
Treonina	Tre	t	101,1051
Triptofano	Trp	w	186,2133
Tirosina	Tir	y	163,1760
Valina	Val	v	99,1326

Fonte: O autor baseado em Peptide 2.0 <www.peptide2.com>.

Também foram consideradas as modificações pós-traducionais no cálculo dos pesos moleculares, como indicado na Tabela 3. As modificações pós-traducionais consideradas, de acordo com Yutin et al. (2013) para proteínas ribossomais, são a acetilação, glutamato e metilação, sendo a acetilação para as proteínas ribossomais S5, S18 e L12; o glutamato para a proteína ribossomal S6; e a metilação para as proteínas ribossomais S11, S12, L3, L7, L11, L16 e L33. Segundo Arnold e Reilly (2015), também há a redução de 131,1963 Da sempre que houver a presença de glicina, alanina, prolina, serina, treonina, valina ou cisteína após a metionina inicial.

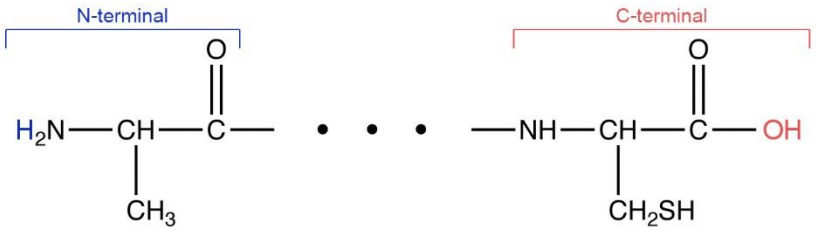
TABELA 3 – Modificações pós-traducionais consideradas pela calculadora de peptídeos

Proteína ribossomal	**Modificação**	**Peso adicionado em Da**
S5, S18, L12	Acetilação	42,0373
S6	Adição de um resíduo de Glutamato	129,1155
S11, L3, L7, L16, L33	Metilação	15,0345
S12	Metilação	48,1075
L11	Metilação	45,1035

Fonte: O autor.

As proteínas analisadas pela calculadora de peptídeos, iniciam com peso de 18,0220 Da. Este valor é referente ao peso de uma suposta molécula de água, considerando o hidrogênio (H) presente na posição N-terminal e a hidroxila (OH) presente na posição C-terminal da cadeia polipeptídica, como é mostrado na Figura 6.

FIGURA 6 – Exemplo de cadeia polipeptídica demonstrando as posições N e C-terminal



Legenda: Começo e término de uma cadeia polipeptídica mostrando o hidrogênio (H) inicial e hidroxila final (OH), comumente perdidos durante a ligação de dois peptídeos.
Fonte: O autor.

4.3 BASE DE DADOS PRESUMIDA

Os pesos moleculares das proteínas calculadas na etapa anterior (item 4.2) formaram uma base de dados, juntamente com suas respectivas taxonomias e nomes das proteínas ribossomais. As proteínas ribossomais escolhidas formaram 60 atributos nesta base de dados, e a taxonomia do microrganismo compôs o último atributo alvo (Tabela 4). Essa base foi chamada de “base de dados presumida”. Na base de dados presumida, os pesos moleculares caracterizam dados do tipo quantitativo contínuo, e as taxonomias descritas são dados do tipo qualitativo nominal.

TABELA 4 – Exemplo de um espectro virtual da base de dados presumida

L1	**L2**	**L3**	**L4**	**L5**	**L6**	**L7**
24447.31	29992.54	22641.04	22337.64	20133.56	19450.29	?
L7a	**L7ae**	**L7/L12**	**L9**	**L10**	**L11**	**L12**
?	10840.4	12279.05	16616.24	17442.33	14900.56	?
L13	**L14**	**L15**	**L16**	**L17**	**L18**	**L19**
16304.77	13222.36	15526.75	16328.1	14058.18	13153.2	12908.11
L20	**L21**	**L22**	**L23**	**L24**	**L25**	**L27**
13420.95	11198.05	12752.84	11092.86	10909.7	?	10236.76
L28	**L29**	**L30**	**L31**	**L32**	**L33**	**L34**
6908.14	7875.23	6427.57	9544.59	?	6103.17	5494.64
L35	**L36**	**S1**	**S2**	**S3**	**S4**	**S5**
7508.74	?	?	28629.81	24130.83	22867.01	17372.16
S6	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**
11588.13	17859.77	14286.51	14136.37	11480.41	14063.29	15123.6
S13	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**
13545.73	?	10483.07	9718.33	10304.98	9415.09	10561.26
S20	**S21**	**S22**	**S31e**	**Taxonomy**		
9257.62	6768.81	?	?	Abiotrophia defectiva		

Fonte: O autor.

A taxonomia herdada da publicação adquirida do repositório do NCBI, foi generalizada para o nível de espécie. No total, a base presumida é composta de 1.949 gêneros e 6.936 espécies distintas, distribuídas em 28.505 registros. O processo de conversão das sequências de aminoácidos para seus respectivos pesos moleculares, formando a base de dados

presumida, demorou 18 dias e 43 minutos para sua execução em um computador Xeon com 24 núcleos (2 GHz cada) e 16GB de memória RAM, rodando em um sistema operacional Linux Ubuntu 3.13 64 bits. A mesma execução em um computador tradicional, como por exemplo um servidor com 8GB de memória RAM e processador Intel i5 com aproximadamente 2 GHz para cada um dos 5 núcleos, demoraria um equivalente a 90 dias de execução.

4.3.1 Pré-processamento da base de dados presumida

Seguindo os conceitos de pré-processamento realizados no aprendizado de máquina, apresentados por Faceli et al. (2011), foram verificadas as características da base de dados presumida. As verificações destas etapas de pré-processamento visaram melhorar a qualidade dos dados, analisando os dados obtidos, dados faltantes, correlações, ruídos e dimensionalidade, elevando assim a capacidade preditiva do modelo.

Sabido que a base de dados presumida foi construída de acordo com as proteínas ribossomais comumente presentes em procariotos (Yutin et al., 2012), os atributos designados já estavam em conformidade com o que se pretendia obter no modelo final do aprendizado de máquina. Assim, não foi realizada nenhuma tarefa de exclusão manual de atributos, integração de atributos ou reamostragem automática dos dados através de algoritmos. Objetos relativos a taxonomias de filo, família e ordem foram retirados, assim como registros de bactérias "*unclassified*", "*uncultured*", "*unidentified*" e "*candidatus*".

Os dados obtidos são considerados desbalanceados devido a variação no número de registros de estirpes adquiridas do repositório. As cinco espécies mais populosas na base de dados presumida são a *Staphylococcus aureus*, *Escherichia coli*, *Mycobacterium tuberculosis*, *Salmonella enterica* e *Acinetobacter baumannii*, apresentando respectivamente 4064, 1955, 1602, 1001 e 665 registros de estirpes. Sendo que 4731 espécies possuem registros únicos, representando aproximadamente 16% dos registros e 63% das espécies. Calculou-se que a média de registros por espécie é de 4, e a média ponderada dos registros da base apresenta 1 registro por espécie.

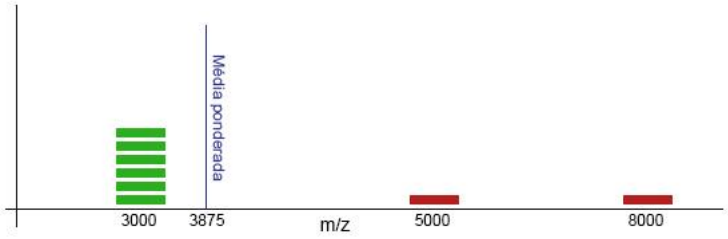
De fato, os números apresentados vão na contramão das recomendações sobre um viés de representatividade para o aprendizado de máquina. Porém, os valores obtidos que compõem a base de dados presumida, são a representatividade atual do conhecimento público acessível sobre microrganismos até então estudados, sequenciados e publicados.

Visando não perder as informações dos dados obtidos, tanto nas taxonomias com grande representatividade, quanto nas taxonomias com pouca representatividade, optou-se por não realizar a redefinição do tamanho dos conjuntos ou pela indução de objetos virtuais. No entanto, com intuito de compensar as classes desbalanceadas e otimizar a busca de intervalos no modelo, adotou-se uma medida constante (de até 1 *m/z*) para os desvios padrões estabelecidos no modelo. Esta medida afeta a busca dos intervalos. Como consequência, as condicionais trazidas para o cálculo da probabilidade são apenas aquelas mais próximas dos *kernels* do modelo. Buscou-se evitar assim um produtório extenso de variáveis, ainda que pequenas, porém presentes nas classes que eram mais populadas. Também foi associado um número pequeno referente às massas moleculares faltantes nas condicionais. Mais detalhes sobre a restrição do desvio padrão e massas moleculares faltantes, encontram-se nas seções 5.1 e 5.2 respectivamente.

Não foram encontrados dados inconsistentes, redundantes, ou que necessitassem de algum tipo de transformação. Como pretendeu-se trabalhar com todas as proteínas ribossomais comumente presentes nos microrganismos, não foi realizada nenhuma tarefa de redução de dimensionalidade da base de dados presumida.

Algumas bactérias do banco apresentam proteínas parálogas, ou seja, proteínas com sequências de aminoácidos e estruturas semelhantes, porém com funções diferentes. Nestes casos, quando uma mesma proteína ribossomal em um mesmo microrganismo apresentou três ou mais registros no NCBI que geraram pesos moleculares diferentes, a média ponderada foi calculada, e o peso mais próximo da média foi escolhido. Assim, escolheu-se os valores mais prováveis, baseado na frequência dos parálogos (Figura 7).

FIGURA 7 – Exemplo de escolha do valor de proteínas parálogas através da média ponderada



Legenda: Os valores incidentes no valor 3000 fazem com que a média se aproxime do valor mais reincidente, sendo este posteriormente escolhido para definir o valor final.
Fonte: O autor.

Quando apenas dois diferentes pesos moleculares foram encontrados, a equação da função da densidade da probabilidade (Fórmula 1) foi utilizada para escolher o peso mais provável. Para possibilitar o cálculo através da função da densidade da probabilidade, a média e desvio padrão de cada gênero foi calculado.

$$f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (1)$$

Na Fórmula 1 o valor de x representa o valor a ser testado, a variável μ representa a média previamente calculada, e σ^2 o valor do desvio padrão. Os ajustes das proteínas ribossomais parálogas foram realizados através de *script*, o qual demorou 20h e 15m para ser executado.

Atualmente, a maior parte dos espectros de massa chega 10 kDa. Há máquinas, como MALDI-TOF, que podem alcançar até 40 kDa. Por essa razão, proteínas com até 40 kDa foram utilizadas. Essa estratégia, também preveniu *outliers* no algoritmo de aprendizado de máquina, causado por grandes sequências de aminoácidos publicadas.

4.4 CONSTRUÇÃO DO MODELO

Quando necessita-se utilizar um aprendizado de máquina supervisionado, também é necessário haver uma generalização nos dados, possuindo vários registros que contribuem como conhecimento para a definição de um mesmo atributo-alvo. Tal generalização foi realizada colocando todas as taxonomias presentes na base de dados presumida, para o nível de espécie. As taxonomias que estavam presentes apenas em nível de gênero mantiveram-se neste mesmo nível taxonômico.

Para construir o modelo foi utilizado o *software* Weka (FRANK, HALL e WITTEN, 2016), o qual é disponibilizado em código aberto com implementação disponível em códigos Java. O algoritmo escolhido para a construção do modelo foi o *Naïve Bayes* com opção *kernels estimator* (JOHN e LANGLEY, 1995). Este algoritmo gera um modelo para classificação, fornecendo um desvio padrão e um grupo de médias (*kernels*) para cada proteína ribossomal de cada espécie fornecida.

Foi criado assim um novo projeto na plataforma *Eclipse*, utilizando a linguagem *Java* e o implementador de bibliotecas *Maven*, para implementar os algoritmos do Weka (FRANK,

HALL e WITTEN, 2016), e realizar o treinamento do aprendizado de máquina. Como base de treinamento foi utilizado a base de dados presumida, gerando um modelo final em formato ".txt". Posteriormente o modelo final foi importado através de script, para alimentar a base de dados do modelo do aprendizado (Tabela 5).

TABELA 5 – Estrutura da base de dados do modelo do aprendizado

Nome do Campo	**Tipo de Dado**
cod_rule	int(11)
specie	varchar(150)
protein	varchar(7)
mean	double
std_dev	double
virtual_std_dev	double
weight_sum	double
rule_precision	double
num_kernels	int(11)

Fonte: O autor.

Sendo "*cod_rule*" a numeração padrão do registro no banco de dados; "*specie*" a taxonomia da espécie generalizada; "*protein*" a proteína ribossomal referente; "*mean*" o valor da média calculada pelo aprendizado de máquina; "*std_dev*" o desvio padrão calculado pelo aprendizado de máquina para o padrão reconhecido; "*virtual_std_dev*" herdado o mesmo valor do *std_dev*, porém com a limitação estipulada; "*weight_sum*" valor de referência para a quantidade de registros utilizados para gerar o padrão; "*rule_precision*" precisão da regra proveniente do modelo gerado; "*num_kernels*" número de *kernels* o qual o registro participa. A base de dados com o modelo do aprendizado, foi então utilizada no software para realizar a classificação. Posteriormente uma modificação foi realizada na base de conhecimento atualizando o campo "*virtual_std_dev*", colocando todos os valores maiores que 1 (um), sendo igual a 1 (um). Mais informações sobre essa alteração será discutida na seção 5.1.

4.5 SOFTWARE DE CLASSIFICAÇÃO - RIBOPEAKS

Para o uso do software é necessário informar as massas moleculares das proteínas ribossomais, de uma bactéria. As massas moleculares devem estar separadas pelo caractere "*;*" (Ex: 24735.79;30140.95;21554…).

Depois de enviar as massas moleculares para a análise, o programa capta cada massa fornecida e tenta assimilar com o modelo criado anteriormente na seção 4.4. Para achar um registro no modelo relacionada com a massa molecular informada, é utilizada a Fórmula 2.

$$f(p) = (p - (\mu - 3\sigma) \geq 0) \cap (p\ - (\mu\ + 3\sigma) \leq 0) \quad (2)$$

Na Fórmula 2, "p" significa o valor da massa molecular informada, "μ" o valor da média do *kernel* e "σ" o valor do "*virtual_std_dev*" do modelo. A Fórmula 2 segue a ideia da distribuição normal, onde é esperado que aproximadamente 99,73% dos dados estejam no alcance de três desvios padrão (ROSS, 2004). Para cada resultado encontrado, foi criado um mapa na memória, onde os resultados de possibilidades de cada massa molecular foram armazenados. Quando houve uma segunda possível massa molecular informada, para uma mesma proteína ribossomal de um mesmo microrganismo no mapa de possibilidades, foi então utilizado a Fórmula 1 para determinar a massa mais provável.

Depois de construir todas as possibilidades em memória, foi então calculada a probabilidade de cada possibilidade baseado no algoritmo do *Naïve Bayes* com *kernels estimator* (JOHN e LANGLEY, 1995), como demonstrado na Fórmula 3.

$$P(o|p) = \frac{P(p|o)\,P(o)}{P(p)} \quad (3)$$

Em adição, para as sessenta possíveis proteínas ribossomais no modelo, para cada pico não presente no mapa, é atribuído um valor único de 6,049268112978591E-6. Esse valor foi considerado representativo para picos faltantes, ele foi baseado na probabilidade de uma massa molecular presente em 0 m/z, dado a média de 40 kDa, e desvio padrão de 40 kDa.

O resultado da análise é apresentado na forma de uma lista, ordenada do mais provável microrganismo até o menos provável. Então, o *script* do software *online* informará os dez microrganismos mais prováveis como retorno para o usuário. Juntamente com a probabilidade do microrganismo, chamada de *Score*, são também informadas três outras medidas: contribuição, paridade parcial e paridade total.

4.5.1 Medida de contribuição

A medida de contribuição informa o quanto cada massa molecular contribuiu para o *Score* (Fórmula 1). Abaixo de cada massa molecular informada pelo usuário e reconhecida em

algum modelo, é informada a probabilidade da densidade da mesma. A condicional desta probabilidade junto com a probabilidade a *priori* da classe formam o *Score* como mostrado na Fórmula 3.

4.5.2 Paridade parcial

A paridade parcial, é o quão próximo os picos informados são das suas respectivas médias encontradas no modelo (Fórmula 4).

$$f(p_1 \dots p_n,\ \mu) = \frac{\left(\left(\sum_{p_1}^{p_n} \mu - \sum_{p_1}^{p_n} (\mu - p)^{-1}\right) * 100\right)}{\sum_{p_1}^{p_n} \mu} \quad (4)$$

Onde "p" é cada massa molecular informada e encontrada no modelo e "μ" a média do *kernel* encontrado no intervalo do modelo.

4.5.3 Paridade total

A paridade total demonstra a cobertura que as massas moleculares obtiveram sobre o modelo da espécie cogitada, considerando a paridade parcial da hipótese (Fórmula 5).

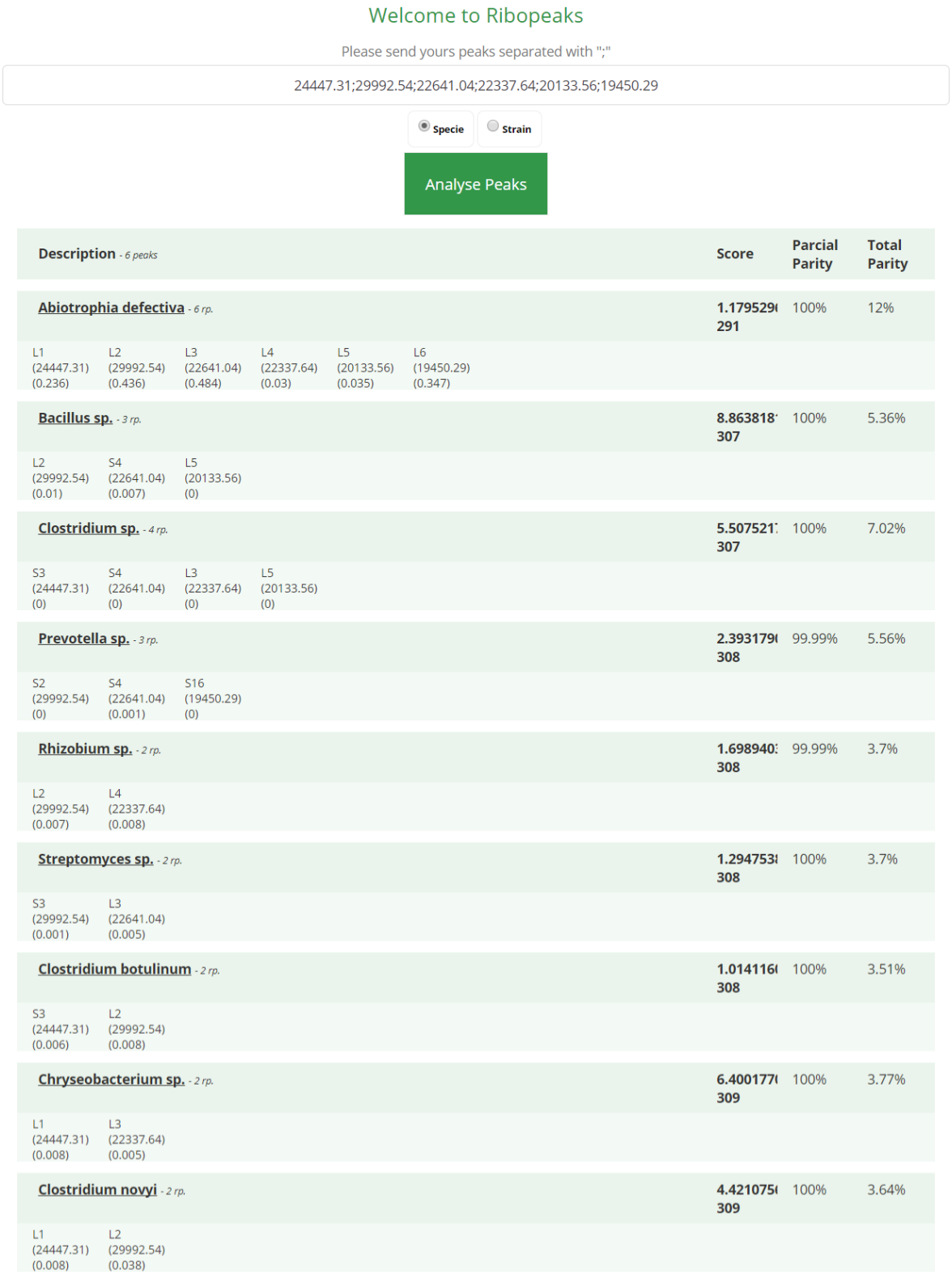
$$f(i,j) = \frac{\left(\left(\frac{i * 100}{j}\right) * P\right)}{100} \quad (5)$$

Onde "*i*" representa o número de massas moleculares encontradas no modelo da espécie cogitada, "*j*" representa o número de massas moleculares totais compreendidas pelo modelo, e "*P*" a paridade parcial obtida pela hipótese.

4.5.4 Interface

Na Figura 8 é mostrado um resultado de classificação de uma lista de massas ribossomais informadas no programa. A lista de resultados traz os melhores resultados e seus respectivos escores, assim como suas medidas de contribuição, paridade parcial e total.

FIGURA 8 – Resultado de uma classificação no *software* Ribopeaks

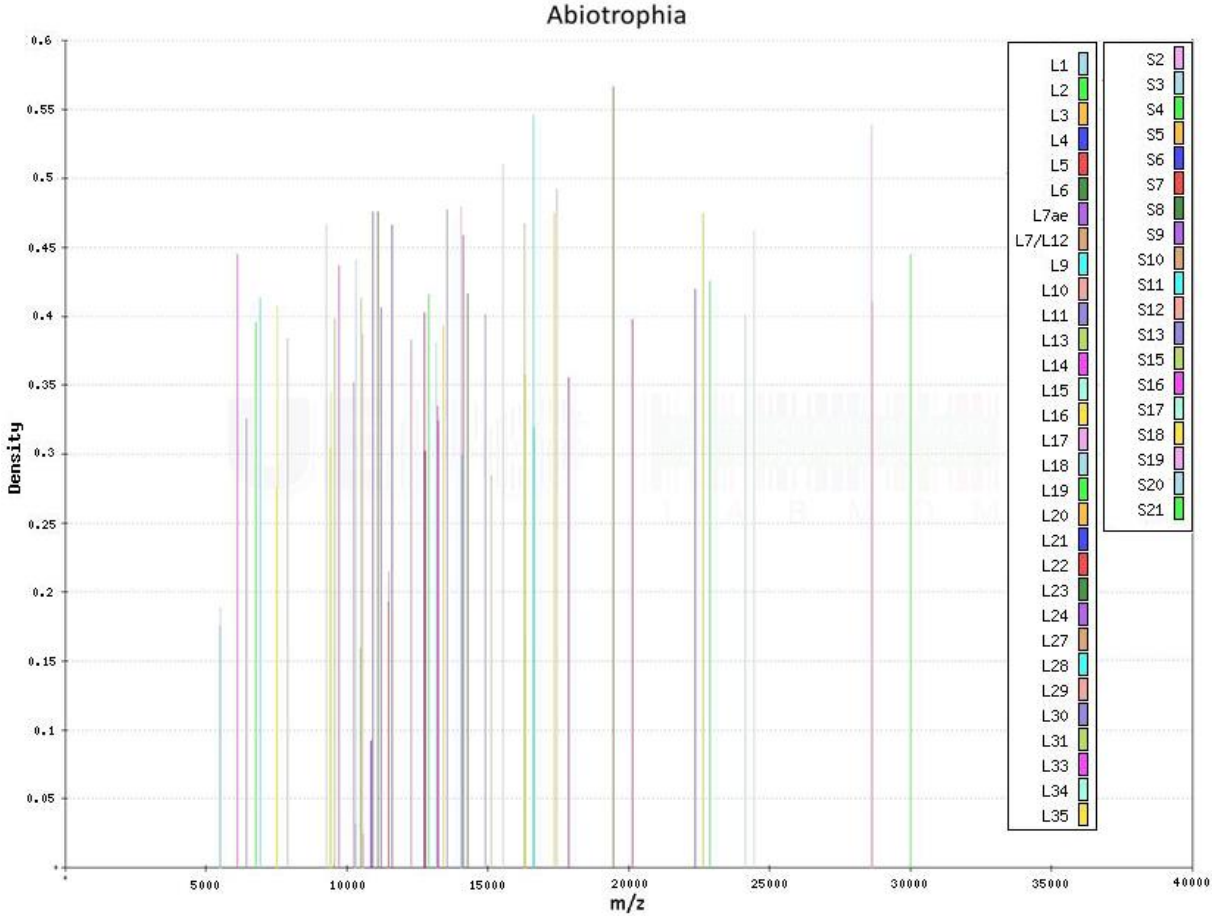


Legenda: Resultado de uma classificação com nove dos dez melhores resultados apresentados no *software* Ribopeaks. Cada uma das classificações trás os valores informados com as prováveis proteínas ribossomais associadas, juntamente com as medidas de *score*, paridade parcial, paridade total e medidas de contribuições.
Fonte: O autor.

4.6 PERFIL DE ESPECTROS DOS GÊNEROS

A base de dados presumida também foi generalizada para o nível taxonômico de gênero, e em sequência executado um novo aprendizado de máquina. O intuito deste modelo gerado, é sua utilização para criar imagens de espectros de massa das bactérias adquiridas. Para esse aprendizado também foi utilizado o algoritmo do *Naïve Bayes*, porém sem opção *kernels estimator* (JOHN e LANGLEY, 1995). Assim, com uma única média, é possível ter uma melhor visualização da curva de distribuição de todos os dados, quando criado o espectro de visualização, possibilitando ao usuário identificar os picos dominantes e pouco frequentes em determinados gêneros (Figura 9).

FIGURA 9 – Perfil com proteínas ribossomais de um determinado gênero



Legenda: O espectro apresentado mostra onde cada proteína ribossomal concentra-se, dado o modelo gerado pelo aprendizado de máquina.
Fonte: O autor.

5 RESULTADOS E DISCUSSÃO

No intuito de classificar diferentes espécies, utilizando informações preditas de massas moleculares referentes a proteínas ribossomais, foi gerada uma base de dados presumida. Essa base foi submetida ao aprendizado de máquina que gerou um modelo de classificação, posteriormente utilizado por um *software* classificador.

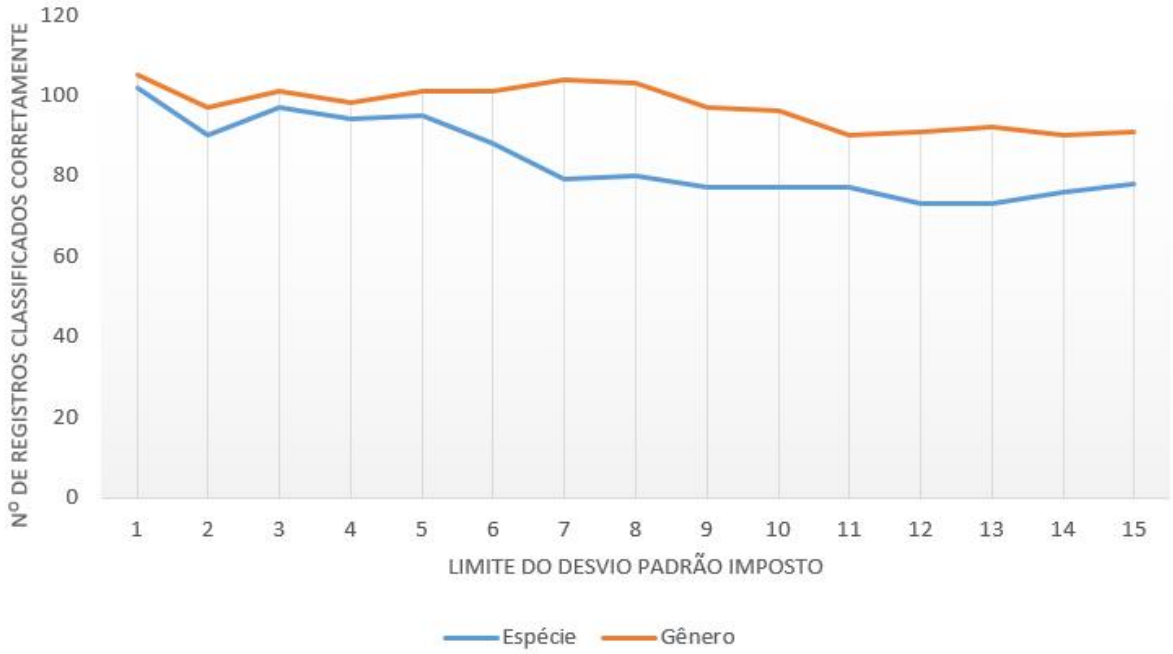
5.1 DESVIO PADRÃO LIMITADO

Como citado na seção 4.4, a estrutura da base de dados do modelo do aprendizado conta com o campo "*virtual_std_dev*". Este campo foi criado como alternativa para limitar o desvio padrão do modelo, sem afetar o dado inicial, posteriormente ainda utilizado para realizar o cálculo da probabilidade condicional. Esta abordagem foi utilizada para reduzir o alcance dos intervalos utilizados para cogitar as possibilidades no mapa de possibilidades, conforme explicado na Fórmula 2. Com isso cogita-se menos possibilidades no mapa de possibilidades, porém se é mais restrito nas informações dos *kernels*. Isto auxilia no desbalanceamento das informações presentes na base dados presumida, pois o produtório não apresentará as condicionais de grandes desvios padrões geradas por vários registros, o que geralmente faz com que o *score* seja mais favorecido a espécies que tenham mais registros em seu aprendizado.

Considerando a utilização de vários núcleos de médias (*kernels*), gerados pelo aprendizado de máquina, observou-se que o sistema de classificação obteve uma melhora em seu desempenho classificatório, sem comprometer o número de proteínas ribossomais encontradas, conforme mostrado na Figura 10.

A distribuição dos dados mostra como algumas proteínas ribossomais possuem uma distribuição concentrada de seus registros em pequenos grupos (Apêndice 1). O resultado da Figura 10 coincide com dados obtidos do teste de classificação de toda a base presumida, testada com limite de 1 e 5. O melhor limite obtido, igual a 1, utilizando a lógica da Fórmula 2, resulta em um total de 3 Da de intervalo para mais e para menos segundo a posição da média. Esse mesmo intervalo já é utilizado na literatura, como mostram Stets et al. (2013).

FIGURA 10 – Teste de classificação de 116 bactérias apresentadas em Ziegler et al. (2015)



Legenda: O gráfico apresenta o resultado dos testes de classificação realizados a nível de gênero (linha laranja), e a nível de espécie (linha azul). O eixo *y* mostra a quantidade de registros testados, e o eixo *x* mostra o limite do desvio padrão.
Fonte: O autor.

Usando o limite do desvio padrão para reduzir o número de possibilidades, obteve-se um ganho de desempenho para analisar as massas moleculares informadas de aproximadamente 94,45%. Dado obtido da análise de 13 massas moleculares, tendo a redução de 2,1 minutos para 7,37 segundos. Valores obtidos de uma média de 15 testes realizados em duas máquinas diferentes com equivalente desempenho.

5.2 DADOS FALTANTES

Muitos valores de proteínas na base de dados presumida, encontram-se faltantes. Isto ocorre devido à falta de uma publicação no repositório do NCBI relativa àquela proteína, ou ao fato de que o microrganismo em questão não possui tal polipeptídio.

O algoritmo *Naïve Bayes* (JOHN e LANGLEY, 1995) é capaz de utilizar uma base de conhecimento com dados faltantes, para realizar o aprendizado de máquina. Mesmo casos em que há muitas proteínas ribossomais faltantes, o registro foi mantido, uma vez que a informação

contida nele pode ser relevante para discernimento no modelo, ou ser referência complementar para outros registros que possuam a informação deste como faltante.

A base de dados presumida possui também desequilíbrios com relação à quantidade de amostras por microrganismos. Algumas espécies são dotadas de uma quantidade grande de exemplares de estirpes, provindas do repositório. No entanto, outras apenas possuem um único registro em relação à sua espécie. Isso acontece devido ao viés de seleção apresentado pelo repositório, e também à falta de dados sobre a real diversidade do Domínio Bacteria e Archaea. Com isso optou-se por manter as informações presentes, pois reproduzem o conhecimento que se tem até o momento, ainda que possam não representar o conhecimento natural em sua totalidade.

Com a limitação definida em 40 kDa (seção 4.3), observou-se que a proteína ribossomal S1 teve perda de aproximadamente 94% em seus dados. Isto sugere que a proteína ribossomal possui grande cadeia polipeptídica, e que sua identificação só pode ser realizada em espectros que abranjam um alcance de m/z de aproximadamente 61 kDa, conforme média calculada. Ainda foram subtraídos da base de dados presumida, registros os quais possuíam apenas de um a três massas moleculares. Um levantamento da quantidade das proteínas ribossomais faltantes é mostrado no Apêndice 2.

5.3 CORRELAÇÃO ENTRE AS PROTEÍNAS RIBOSSOMAIS

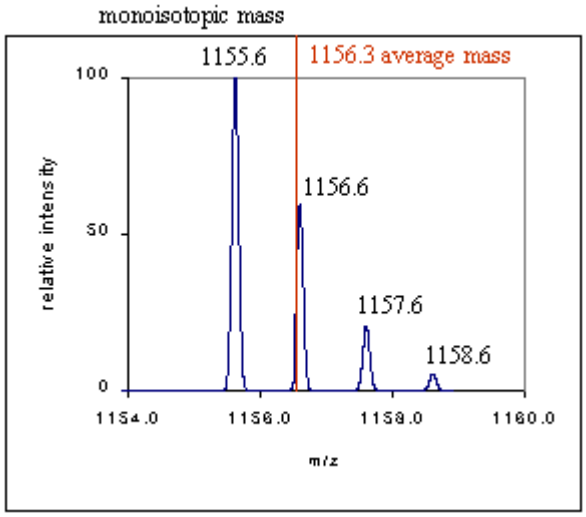
A correlação das proteínas ribossomais na base de dados presumida foi testada. Como mostra o Apêndice 3, nenhuma correlação que ultrapassasse um limiar de 0,7 foi encontrada. Sugerindo e suportando a independência das proteínas para identificação de seus respectivos microrganismos, ideia em que se baseia a metodologia de aprendizado de máquina *Naïve Bayes* (JOHN e LANGLEY, 1995).

5.4 PESOS ISOTÓPICOS

A seção 4.2, referente a criação da calculadora de peptídeos, esclarece a utilização dos pesos isotópicos médios estimados. Estes pesos podem ser assumidos de duas formas: pesos monoisotópicos e pesos isotópicos médios estimados.

O peso monoisotópico é a massa do pico isotópico, o qual a composição elemental é composta do mais abundante desse elemento. Já o peso isotópico médio estimado é calculado pela média dos picos de massas obtidas com suas abundâncias (http://www.ionsource.com/tutorial/isotopes/slide8.htm). A Figura 11 mostra a diferença dos dois pesos, de acordo com o composto C_{48} H_{82} N_{16} O_{17} (Poly-Alanine).

FIGURA 11 – Diferença do peso monoisotópico e isotópico médio estimado



Legenda: O exemplo mostra como a escolha do mais abundante peso isotópico, pode representar uma extremidade da realidade da distribuição dos pesos. Isto pode não representar o real peso do peptídeo, ou haver falta de acurácia quando comparado com pesos reais.
Fonte: Ion Source <http://www.ionsource.com/tutorial/isotopes/slide8.htm>.

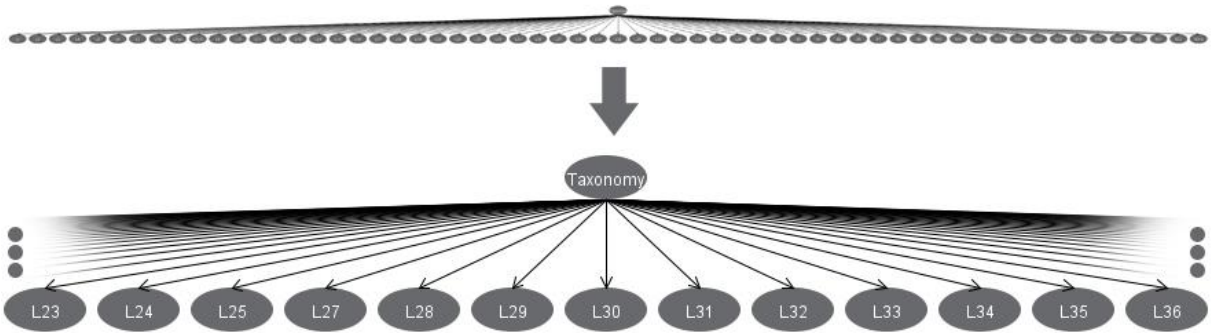
5.5 ANÁLISE DO ESPECTRO DE PROTEÍNAS RIBOSSOMAIS UTILIZANDO APRENDIZADO DE MÁQUINA

Baseado nos trabalhos científicos apresentados, escolheu-se a metodologia de predição de pesos de massas moleculares para criação de uma base de dados presumida, como mostrado na seção 4. Também foi considerado os pontos levantados por Tamura, Hotta e Sato (2013), utilizando-se proteínas ribossomais, que foram obtidas de dados públicos já sequenciados e publicados por autores em um repositório (seção 4.1). Massas moleculares foram calculadas analisando suas probabilidades e eliminando falsos positivos (seção 4.2 e 4.3). Também foram ponderadas as modificações pós-traducionais, como mostra a seção 4.2.

Analisando a estrutura dos dados, buscou-se um algoritmo de aprendizado de máquina supervisionado capaz de resultar em um modelo com valores que determinassem intervalos e qualidades para os intervalos obtidos, assegurando assim a intensidade dos mesmos com o modelo da classificação. Deste modo, dados muito esparsos que não têm um conhecimento bem definido, são penalizados na classificação.

A base de dados presumida foi testada no algoritmo de aprendizado *BayesNet*, no *software* Weka (FRANK, HALL e WITTEN, 2016). O próprio algoritmo é responsável por testar e avaliar as interdependências dos atributos para gerar a classificação, disponibilizando inclusive um gráfico que mostra a relação dos atributos com o atributo meta. No teste realizado todos os atributos, ou seja, todas as proteínas ribossomais, estão independente e diretamente ligadas ao atributo meta, neste caso a taxonomia das bactérias (Figura 12).

FIGURA 12 – Todas as proteínas ribossomais ligadas de forma independente com a taxonomia



Legenda: Grafo gerado pelo algoritmo de aprendizado de máquina da Rede Bayesiana, mostrando a relação direta de cada atributo diretamente com o atributo alvo "*Taxonomy*".
Fonte: O autor.

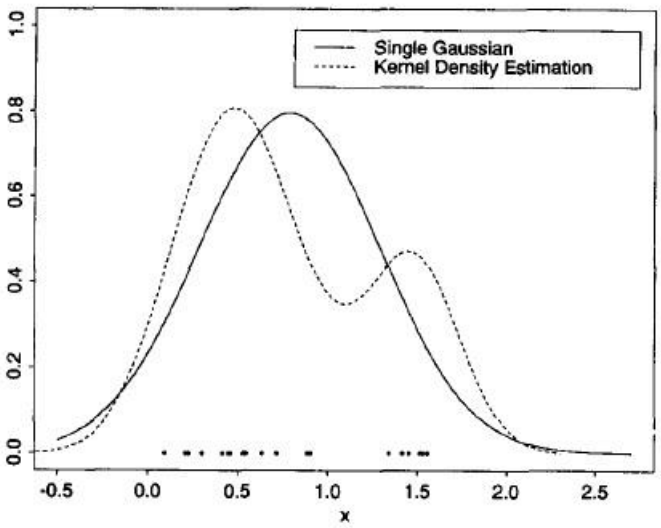
Mostrando a independência dos atributos para com o atributo meta, mas ainda usando um classificador probabilístico, foi testado o algoritmo *Naïve Bayes* (JOHN e LANGLEY, 1995). Com este algoritmo de aprendizado é possível obter um modelo que contém médias e desvios padrões para cada proteína ribossomal, de cada espécie informada. Com isso é possível estipular os intervalos da distribuição dos dados, como também qualificar os mesmos.

Para realizar a classificação neste trabalho, pretendeu-se encontrar também as possíveis proteínas ribossomais dado uma sequência de massas moleculares. Para isso, foi

utilizado os desvios padrões do modelo como referência na definição das hipóteses das proteínas ribossomais possíveis.

O mais adequado algoritmo encontrado para realizar esta definição, e posteriormente a classificação em si, foi o *Naïve Bayes* com opção *kernels estimator* (JOHN e LANGLEY, 1995). A opção *kernels estimator* permite que *n* grupos de médias sejam gerados na criação do modelo, proporcionando uma maior facilidade e confiança em encontrar a qual(is) registro(s) do modelo uma massa molecular pode ser associada. A Figura 13 mostra a diferença entre o modelo do *Naïve Bayes* com e sem *kernels*, na distribuição dos dados.

FIGURA 13 – O efeito do uso de uma única média contra o método com *kernels* para estimar a densidade de uma variável contínua



Legenda: Comparação das curvas de distribuição dos dados entre uma distribuição gaussiana com uma média, e a distribuição gaussiana formada com a denominação de mais médias, ou *kernels*.
Fonte: John e Langley, 1995.

5.6 RESULTADOS DA CLASSIFICAÇÃO DOS MICRORGANISMOS DA BASE DE DADOS PRESUMIDA

Todos os registros da base dados presumida foram posteriormente testados com o software de classificação. Em um total de 28.505 registros analisados, o software foi capaz de classificar corretamente a nível de espécie 27.032 registros (94.83%), considerando as espécies correlatas (Apêndice 4). Outros 1.465 registros (5,14%) foram classificados nas posições 2 a

10 da lista de classificação, tendo apenas 8 registros (0,03%) classificados erroneamente. Também foi observado o acerto a nível taxonômico de gênero, que obteve 28.131 acertos diretos (98.69%), 366 acertos (1,28%) nas posições 2 a 10, e 8 registros (0,03%) erroneamente classificados (Tabela 6).

TABELA 6 – Resultados da classificação da base de dados presumida

Nível Taxonômico	**Nº de registros**	**Acertos**	**Acertos em posição 2 a 10**	**Erros**
Espécie	28.505	27.032* (94,83%)	1.465 (5,14%)	3 (0,03%)
Gênero	28.505	28.131 (98,69%)	366 (1,28%)	3 (0,03%)

* considerando as espécies correlatas.

Fonte: O autor.

De um total de 6.936 espécies distintas informadas no aprendizado de máquina, 99 espécies não foram classificadas corretamente. Entretanto, todos esses microrganismos classificados erroneamente apresentavam alta proximidade filogenética com o resultado da classificação, quando comparado os seus genes 16S rRNA. Um exemplo é a classificação errônea da espécie *Escherichia coli*, identificada com *Shigella boydii* (ZUO; XU e HAO, 2013). Estas espécies são muito próximas filogeneticamente, sendo somente diferenciáveis através do alinhamento completo de seus genomas ou por outros métodos moleculares.

Outro caso pode ser observado na classificação da *Staphylococcus aureus*, classificada como *Staphylococcus argentus*. Como reportado na literatura (CHANTRATITA et al., 2016), estas espécies são comumente identificadas de forma errônea, pois muitas de suas propriedades fenotípicas se assemelham. Isto inclui seus padrões ribossomais, que são geralmente utilizados para identificação, como ocorre com o 16S rRNA. A saída para diferenciação de tais espécies exige uma análise de outros genes, como é feito na técnica de MLST (*Multilocus Sequence Typing*, Sequenciamento de *Multilocus*).

Em todos os casos de erros na classificação dos microrganismos da base de dados presumida (a base de dados sendo comparada contra ela mesma), observou-se em média 95% de similaridade genômica entre as espécies submetidas para a classificação (*query*) e a espécie de maior escore presente no modelo (*subject*). Todos os casos e referências podem ser encontrados no Apêndice 4.

Algumas classificações obtiveram o nome de microrganismos que foram reclassificados posteriormente. Como exemplo a *Agrobacterium tumefaciens*, classificada como *Agrobacterium fabrum*. Esta espécie foi analisada e reclassificada em 2001 (WOOD et al,. 2001), porém ambos os nomes permanecem no repositório do NCBI.

Outro problema observado nas classificações foi decorrente das taxonomias dos registros adquiridos do repositório. Um exemplo desse tipo de ocorrência foi a *Vibrio **c**holerae* e *Vibrio **C**holerae*, ou *Bordetella holm**esii*** e *Bordetella holm**seii***, treinadas como bactérias diferentes, porém classificadas entre si. Outro exemplo foi a espécie *Vibrio fischeri*, classificada como *Aliivibrio fischeri*. Também a *Burkholderia pseudomallei* e *Burkholderia mallei*; ou *Bacillus licheniformis* e *Bacillus paralicheniformis*. Esses casos ainda obtiveram classificação próxima das suas taxonomias originais, ficando na maioria das vezes em segunda posição na lista de classificação.

5.7 CLASSIFICAÇÃO DE DADOS BIOLÓGICOS

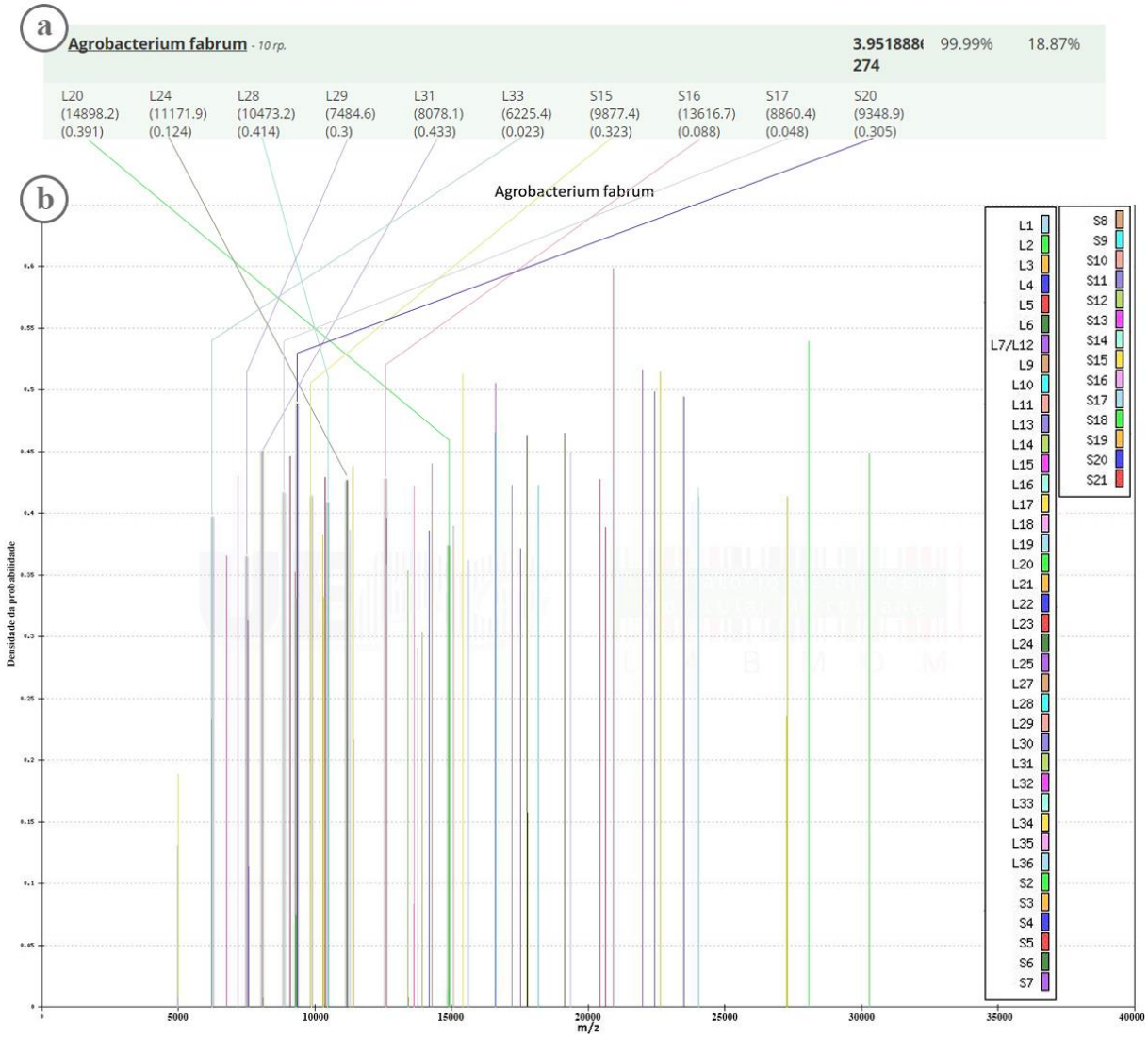
No trabalho de Ziegler et al. (2015), foi demonstrado o uso de proteínas ribossomais para distinção de 116 microrganismos dos gêneros: *Agrobacterium, Azorhizobium, Bradyrhizobium, Burkholderia, Cupriavidus, Ensifer, Mesorhizobium, Methylobacterium, Microvirga* e *Rhizobium*. Estes microrganismos estão associados a fixação de nitrogênio em leguminosas, e representam as principais bactérias do grupo *Rhizobia*.

Os dados foram coletados do material suplementar disponibilizado pela publicação, estes mesmos dados não compuseram a base de dados presumida, tampouco participaram do aprendizado de máquina. Os dados de Ziegler et al. (2015) são provenientes de origem *in vivo*, sendo as 13 proteínas ribossomais selecionadas de acordo com a comparação de espectros reais e virtuais desses 116 microrganismos selecionados. Assim, pretendeu-se nesta análise confrontar os dados gerados *in vivo* com os dados criados *in silico* neste trabalho (Figura 14).

Para estes microrganismos a classificação obteve um acerto direto de 102 registros em nível de espécie, e de 105 registros para acerto em gênero. Obteve-se assim o equivalente a 87,93% para acertos em espécies, e 90,51% de acerto em nível de gênero (Tabela 7). Houveram 4 registros da espécie *Bradyrhizobium canariense*, que foram considerados por não haver um registro correspondente no modelo. Estas mesmas tiveram seu acerto em nível de gênero, e

nenhuma outra espécie de *Bradyrhizobium* foi classificada em seus testes. Aparentemente não há outra espécie no conhecimento treinado, que apresente relação filogenética mais próxima.

FIGURA 14 – Comparação de um resultado com o modelo gerado pelo aprendizado de máquina



Legenda: Na parte A da Figura 14 é mostrado o resultado da classificação de 13 proteínas ribossomais adquiridas em Ziegler et al. (2015). A parte B da Figura 14 contém o espectro do modelo criado pelo aprendizado de máquina, o eixo *x* representa o peso da massa molecular, e o eixo *y* mostra a medida de contribuição.
Fonte: O autor.

O resultado mostra concordância dos pesos moleculares criados *in silico* com a calculadora de peptídeos, com os valores apresentados pelas 13 proteínas ribossomais no trabalho de Ziegler et al. (2015).

TABELA 7 – Resultados da classificação dos dados apresentados em Ziegler et al. (2015)

Nível Taxonômico	**Nº de registros**	**Acertos**	**Acertos em posição 2 a 10**	**Erros**
Espécie	116	102 (87,93%)	9 (7,75%)	5 (4,31%)
Gênero	116	105 (90,51%)	6 (5,17%)	5 (4,31%)

Fonte: O autor.

Espectros biológicos de MALDI-TOF pertencentes ao banco do LABMOM (Laboratório de Biologia Molecular Microbiana) da Universidade Estadual de Ponta Grossa, foram também testados. Como esses espectros contém picos correspondentes a proteínas ribossomais e não ribossomais, a classificação taxonômica utilizando o banco gerado nesse trabalho se mostrou menos eficientes. Para uma correta classificação, torna-se necessário filtrar as massas moleculares ribossomais, para então testá-las com o modelo gerado. As bactérias do gênero *Pseudomonas* foram as melhores classificadas, principalmente quando a medida de contribuição foi restringida à um mínimo de 0,01.

5.8 CLASSIFICAÇÃO ATRAVÉS DE CHAIN-RULES

No intuito de utilizar dados provenientes diretamente de espectros de MALDI-TOF, foi abordada uma técnica para tentar separar valores de massas moleculares dos espectros, referentes às proteínas ribosomais baseado no modelo de aprendizado criado (seção 4.4). A técnica utilizada é chamada de *Chain-rules*, e consiste na ideia de reduzir o espaço de buscas de possíveis soluções. No entanto, como demonstrado a seguir, a técnica gera um número muito elevado de possibilidades para serem testadas.

Assim, para haver um tempo hábil de espera de processamento por parte do usuário, seria necessário utilizar técnicas de computação paralela para que todas as possibilidades fossem analisadas. A técnica foi programada e parcialmente testada, considerando que os testes foram realizados com apenas 4 valores de massas moleculares pertencentes a proteínas ribossomais e 3 valores não pertencentes. Estes testes demonstraram uma classificação correta, porém gastaram mais de 24h para efetuar a análise. Em geral, a lista de massas moleculares de polipeptídios obtidos de espectros de massa, variam na média de 70 valores, no alcance de 0 a

20 kDa. O viés de seleção dos testes também representou uma quantidade muito pequena da amostra da base de dados presumida, não podendo assim afirmar a efetividade da técnica.

A técnica consiste em presumir as possibilidades totais das massas moleculares informadas, pertencerem às proteínas ribossomais definidas na base de dados presumida, calculando suas probabilidades através do modelo. Este espaço de possibilidades foi reduzido através da combinação em pares das possibilidades cogitadas.

Para realizar a programação da técnica, a média e o desvio padrão de cada proteína ribossomal *a priori* da base de dados presumida foi calculada. Utilizando-se destes valores, dado uma lista de massas moleculares informada, cogitou-se as possibilidades iniciais de cada assimilação reconhecida nos intervalos (Fórmula 6).

$$f(m) = (m - (\mu - 3\sigma) \geq 0) \cap (m\ - (\mu\ + 3\sigma) \leq 0) \quad (6)$$

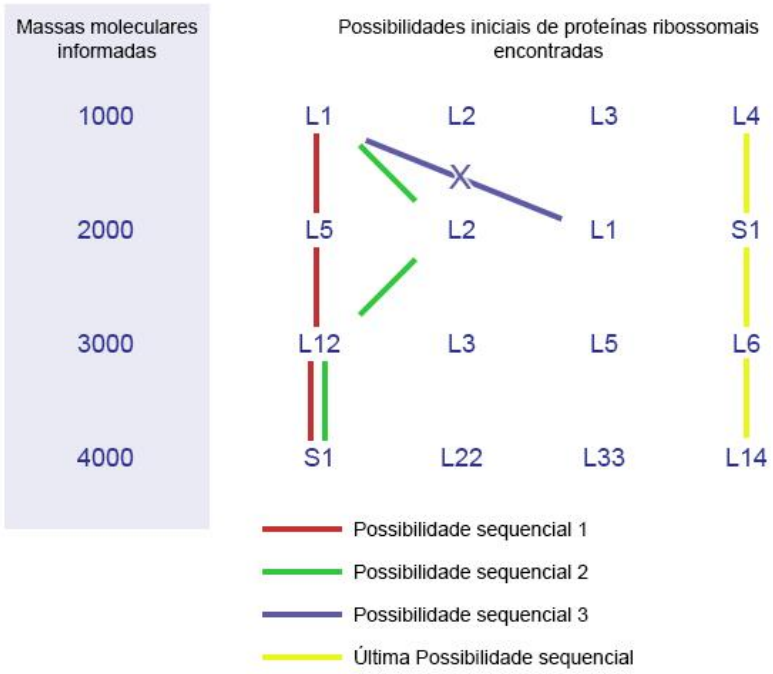
Na Fórmula 6, "*m*" significa o valor da massa molecular informada, "μ" o valor da média a priori da proteína ribossomal e "σ" o valor do desvio padrão calculado. Assim, cada possibilidade inicial representa uma massa molecular associada a uma proteína ribossomal, porém uma mesma massa molecular pode estar presente em *n* possibilidades iniciais, associada a diferentes proteínas ribossomais.

Após cogitar as possibilidades iniciais de cada massa molecular, foram organizadas as possibilidades sequenciais, de acordo com as massas moleculares informadas. Assim, uma lista de valores com *x* massas moleculares, oferecem uma quantidade de x^p, onde *p* representa a quantidade de proteínas ribossomais possíveis. Desta forma as possibilidades formam uma árvore de possibilidades, como demonstrado na Figura 15.

Para poupar memória e desempenho, casos onde diferentes massas moleculares estão associadas a mesma proteína ribossomal, como a "Possibilidade sequencial 3" da Figura 15, não foram assimilados. Realizando esta poda, várias outras possibilidades também foram ignoradas, uma vez que uma poda prematura ignora posteriores sequências que seriam realizadas nos nós abaixo do nível podado.

Após definido as possibilidades sequenciais, foram definidas as combinações em pares das possibilidades sequenciais. As combinações em pares das possibilidades sequenciais foram feitas percorrendo recursivamente as possibilidades sequenciais e concatenando suas hipóteses. Foram descartadas situações, onde, na combinação do par houve concorrência de diferentes massas moleculares para definição de uma mesma proteína ribossomal.

FIGURA 15 – Árvore de possibilidades sequenciais



Legenda: Para cada massa molecular informada e suas correspondentes possibilidades, o esquema mostra a lógica do algoritmo para criação das possibilidades sequenciais, realizando uma busca em profundidade pelas combinações, ignorando equiparidade de informações.
Fonte: O autor.

Obtendo as possibilidades sequenciais e as combinações em pares, foi testado para cada microrganismo do modelo de aprendizado de máquina, a técnica pretendida. Para isso, cada possibilidade sequencial foi condicionada ao microrganismo em questão. Foi encontrado no modelo seus *kernels* mais próximos, e calculado o produtório das suas hipóteses. Em sequência foi realizado a soma das probabilidades de cada possibilidade sequencial condicionada, isso formou o *score* final dentro do espaço de buscas.

Afim de reduzir o espaço de busca das probabilidades calculadas, foi testado também para cada microrganismo do modelo, as combinações em pares das possibilidades sequenciais. Estas também foram condicionadas ao microrganismo em questão, e calculadas através do produtório de hipóteses concatenadas, em conjunto da probabilidade a priori do respectivo microrganismo. Estes valores foram então subtraídos do *score* final previamente calculado.

O *chain-rules* estipula que o espaço de busca deve ser diminuído realizando não somente a combinação dos pares, mas também das trincas, quádruplas, em diante, até que se tenha a combinação total. Como pretendia-se testar a eficiência da técnica, e como a

combinação dos pares representam um produtório menor, portanto maiores valores para decréscimo do *score*, realizou-se apenas até a combinação dos pares.

Devido ao alto número de combinações não possível retirar os resultados desta metodologia. Um teste com cinco massas moleculares foi realizado, apresentando correta classificação, porém demandou mais de 24 horas para ser realizado. Tendo que muitos espectros apresentam mais de 50 proteínas ribossomais, não houve como se testar mais a metodologia sem a necessidade de outras metodologias de otimização e distribuição do processamento.

6 CONCLUSÃO

Neste trabalho investigou-se o uso de um modelo classificatório, construído a partir de um aprendizado de máquina. Utilizou-se como base de treinamento uma base de dados gerada com espectros virtuais de microrganismos obtidos de um repositório público. A base de dados com espectros virtuais foi criada com informações de pesos moleculares estimados de proteínas ribossomais, os pesos moleculares estimados foram calculados através de uma calculadora de peptídeos construída neste mesmo trabalho.

Assim, concluiu-se que os pesos moleculares estimados pela calculadora de peptídeos, juntamente com suas modificações pós-traducionais, obtiveram relação com pesos moleculares já calculados na literatura e apresentados por Ziegler et al. (2015). Isto também mostrou uma válida representatividade das sequências obtidas do repositório, não demonstrando grandes diferenças nos valores das massas moleculares preditas.

A escolha de manter a dimensionalidade da base de dados presumida, possibilitou atribuir as hipóteses de a quais proteínas ribossomais os valores informados poderiam pertencer. Porém executar testes com outros algoritmos de aprendizado de máquina tornou-se desafiador durante a execução do trabalho, uma vez que além da alta dimensionalidade também havia uma alta quantidade de espécies, registros e dados faltantes ou inexistentes.

O algoritmo de aprendizado de máquina *Naïve Bayes* com opção *kernels estimator* (JOHN e LANGLEY, 1995), mostrou-se eficaz em predizer um modelo para classificação do conjunto apresentado. Este foi capaz de compreender mais a distribuição dos dados, oferecer uma alternativa para compensar algumas espécies, calcular um desvio padrão mínimo e trabalhar com dados faltantes ou não existentes na base de dados presumida.

Os resultados obtidos mostraram a capacidade de classificar as espécies treinadas, bem como classificar algumas espécies com proximidades fisiológicas já conhecidas pela literatura. Estes casos coincidem com outras análises utilizando dados ribossomais, como o 16S rRNA. Sugere-se que é possível montar um modelo ainda mais discriminatório tendo mais amostras destes tipos de espécies. Isto é uma vantagem neste tipo de análise pois a diferença entre pesos é dada por números reais, enquanto que pelos 16S rRNA a diferença pode aparecer em um nucleotídeo.

Também foi observado o poder classificatório em espécies em que houve reclassificação taxonômica em registros do repositório. Essas espécies sofreram alteração em

suas taxonomias durantes os anos, mas suas fisiologias foram reconhecidas na classificação. Outros registros que não sofreram reclassificação taxonômica, porém continham divergências nas taxonomias de suas publicações, também foram reconhecidos.

O *software* de classificação desenvolvido foi capaz de cogitar as proteínas ribossomais inseridas, baseando-se no modelo, e classificar em tempo hábil uma lista de massas moleculares fornecidas. Os valores de retorno do *software* de classificação mostram os escores obtidos, bem como a proximidade das informações inseridas com o modelo de referência, através da "Paridade Parcial" e da "Paridade Total". Também é mostrado para cada massa molecular informada, os valores que mais definem o escore final daquela classificação, mostrando assim quais massas moleculares mais destacam-se para identificação da espécie.

A busca pelos intervalos no modelo mostrou-se satisfatória para cogitar as hipóteses das proteínas ribossomais, dada uma lista de massas moleculares contendo apenas outros valores de massas moleculares ribossomais. Nos casos testados de listas de massas moleculares de espectros inteiros, que contêm informações não apenas de proteínas ribossomais, mas também de outros tipos de macromoléculas, os valores não ribossomais confundiram a cogitação das hipóteses. Para isso, seria necessário realizar um novo aprendizado de máquina, ou outro algoritmo discriminativo capaz de, ao menos parcialmente, identificar os possíveis valores pertencentes às proteínas ribossomais. Isto poderia compor um trabalho futuro a ser feito, reestruturando e utilizando a própria base de dados presumida como referência.

A classificação através do *chain-rules* utilizando o modelo gerado mostrou-se promissora, porém com o alto número de possibilidades de combinações não foi possível testar com mais de cinco valores de massas moleculares, nem observar a capacidade do algoritmo em filtrar os valores ribossomais de espectros inteiros. Para executar esta ideia em um trabalho futuro, seria necessário implementar os códigos desenvolvidos em uma estrutura de computação paralela, atentando para o número de possibilidades e de distâncias estipuladas para os desvios padrão.

REFERÊNCIAS

ARNOLD, R.J.; REILLY, J.P. Observation of Escherichia coli Ribosomal Proteins and Their Posttranslational Modifications by Mass Spectrometry. **Analytical biochemistry**, ed. 269, n. 1, pg. 105-112, 1999.

BEER, I. et al. Improving large-scale proteomics by clustering of mass spectrometry data. **Proteomics**, ed. 4, n. 4, pg. 950-960, 2004.

BINZ, P. et al. A molecular scanner to automate proteomic research and to display proteome images. **Analytical chemistry**, ed. 71, n. 21, pg. 4981-4988, 1999.

BRIM, C. E. R. MS-ML Studio: A Generic Free and Open-Source Tool for Mass Spectrometry Analysis and Classification. **Universidade Federal do Paraná**, 2015. (Não publicado)

BULGARELLI, D. et al. Structure and functions of the bacterial microbiota of plants. **Annual review of plant biology**, ed. 64, pg. 807-838, 2013.

CHANTRATITA, N. et al. Comparison of community-onset Staphylococcus argenteus and Staphylococcus aureus sepsis in Thailand: a prospective multicentre observational study. **Clinical microbiology and infection**, ed. 22, n. 5, pg. 458-e11, 2016.

DE BRUYNE, K. et al. Bacterial species identification from MALDI-TOF mass spectra through data analysis and machine learning. **Systematic and applied microbiology** ed. 34, n. 1, pg. 20-29, 2011.

DEMIREV, P. A. et al. Microorganism identification by mass spectrometry and protein database searches. **Analytical chemistry**, ed. 71, n. 14, pg. 2732-2738, 1999.

FRANK, E.; HALL, M. A. E WITTEN, I. H. The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", **Morgan Kaufmann**, ed. 4, 2016.

FACELI, K. et al. Inteligência Artificial: Uma abordagem de aprendizado de máquina. **Rio de Janeiro**: LTC, v. 2, p. 192, 2011.

FOLEY, J. A. et al. Solutions for a cultivated planet. **Nature** ed. 478, n. 7369, pg. 337-342, 2011.

GANS, J.; WOLINSKY, M.; DUNBAR, J. Computational improvements reveal great bacterial diversity and high metal toxicity in soil. **Science** ed. 309, n. 5739, pg. 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 infectologia**, ed. 29, n. 3, pg. 263-272, 2012.

GIBB, S. Entwicklung einer flexiblen bioinformatischen Plattform zur Analyse von Massenspektrometriedaten, **Medizinischen Fakultät der Universität Leipzig**, 2014.

GIBB, S.; STRIMMER, K. MALDIquant: a versatile R package for the analysis of mass spectrometry data. **Bioinformatics**, ed. 28, n. 17, pg. 2270-2271, 2012.

HETTICK, J. M. et al. Discrimination of intact mycobacteria at the strain level: A combined MALDI-TOF MS and biostatistical analysis. **Proteomics**, ed. 6, n. 24, pg. 6416-6425, 2006.

JOHANSSON, P. et al. SPECLUST: a web tool for clustering of mass spectra. **J Proteome Res**, 2007.

JOHN, G. H.; LANGLEY, P. Estimating continuous distributions in Bayesian classifiers. In Proceedings of the Eleventh conference on Uncertainty in artificial intelligence. **Morgan Kaufmann Publishers Inc.**, pg. 338-345, 1995.

MONIGATTI, F.; BERNDT, P. Algorithm for accurate similarity measurements of peptide mass fingerprints and its application. **Journal of the American Society for Mass Spectrometry**, ed. 16, n. 1, pg. 13-21, 2005.

MITCHELL, T. M. Machine learning. **Burr Ridge, IL: McGraw Hill**, ed. 45, n. 37, pg. 870-877, 1997.

MÜLLER, M. et al. Visualization and analysis of molecular scanner peptide mass spectra. **Journal of the American Society for Mass Spectrometry,** ed. 13, n. 3, pg. 221-231, 2002.

RESSOM, H. W. et al. Peak selection from MALDI-TOF mass spectra using ant colony optimization. **Bioinformatics**, ed. 23, n. 5, pg. 619-626, 2007.

ROSS, S. M. Introduction to probability and statistics for engineers and scientists. **Academic press**, 2004.

SANTOS, R. S. MS-Analyser: programa de computador para identificação de microrganismos por análise de MALDI-TOF. **Universidade Federal do Paraná**, 2013. (Não publicado)

SCHLOSS, P.; HANDELSMAN, J. Toward a census of bacteria in soil. **PLoS Comput Biol** ed. 2, n. 7. pg. 92, 2006.

SCHMIDT, F. et al. Iterative data analysis is the key for exhaustive analysis of peptide mass fingerprints from proteins separated by two-dimensional electrophoresis. **Journal of the American Society for Mass Spectrometry**, ed. 14, n. 9, pg. 943-956, 2003.

STELZL, U. et al. Ribosomal proteins: role in ribosomal functions. **eLS**. 2001.

STETS, M.I. et al. Rapid identification of bacterial isolates from wheat roots by high resolution whole cell MALDI-TOF MS analysis. **Journal of biotechnology**, ed. 165, n. 3, pg. 167-174, 2013.

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**, ed. 24, n. 8, pg. 1185-1193, 2013.

TERAMOTO, K. et al. Phylogenetic classification of Pseudomonas putida strains by MALDI-MS using ribosomal subunit proteins as biomarkers. **Analytical chemistry**, ed. 79, n. 22, pg. 8712-8719, 2007.

TIBSHIRANI, R. et al. Sample classification from protein mass spectrometry, by 'peak probability contrasts'. **Bioinformatics,** pg. 20, n. 17, pg. 3034-3044, 2004.

ULINTZ, P. J. et al. Improved classification of mass spectrometry database search results using newer machine learning approaches. **Molecular & Cellular Proteomics**, ed. 5, n. 3, pg. 497-509, 2006.

VESSEY, J. K. Plant growth promoting rhizobacteria as biofertilizers. **Plant and soil** ed. 255. N. 2, pg. 571-586, 2003.

VILLANUEVA, J. et al. Serum peptide profiling by magnetic particle-assisted, automated sample processing and MALDI-TOF mass spectrometry. **Analytical chemistry**, ed. 76, n. 6, pg. 1560-1570, 2004.

WATSON, J. D. et al. Biologia molecular do gene. **Artmed Editora**, 2015.

WILLARD, H. H. et al. **Instrumental methods of analysis**, 1988.

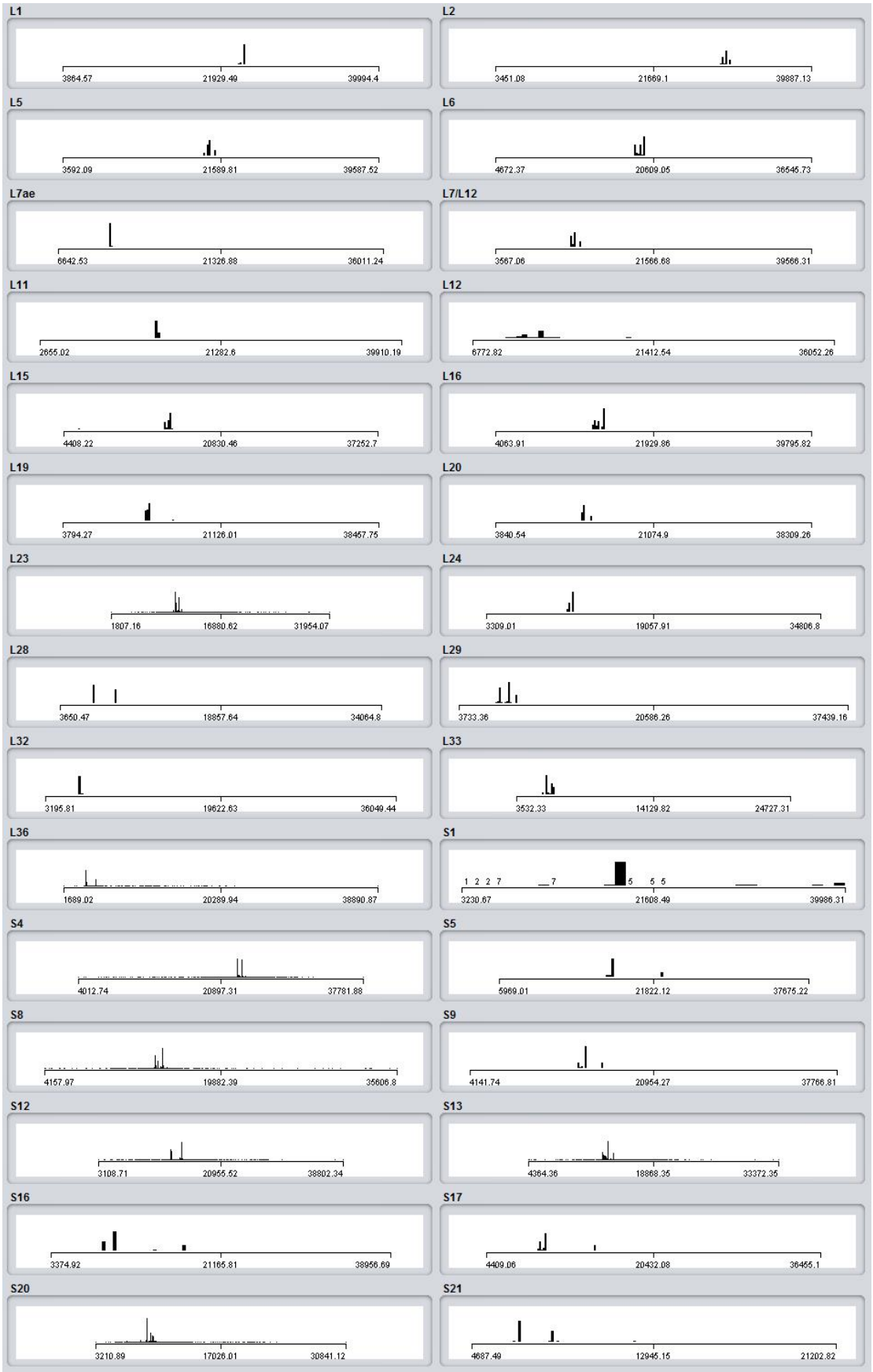
WOOD, D.W. et al. The genome of the natural genetic engineer Agrobacterium tumefaciens C58. **Science**, ed. 294, n. 5550, pg. 2317-2323, 2001.

YUTIN, N. et al. Phylogenomics of prokaryotic ribosomal proteins. **PLoS One**, ed. 7, n. 5, pg. 36972, 2012.

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

ZUO, G.; XU, Z.; HAO, B. Shigella strains are not clones of Escherichia coli but sister species in the genus Escherichia. **Genomics, proteomics & bioinformatics**, ed. 11, n. 1, pg. 61-65, 2013.

APÊNDICE 1 – Distribuição dos dados *a priori* por proteína ribossomal




L3
4600.46
22288.27
39976.09
L4
3207.51
21301.01
39394.5
L7
6412.56
23177.11
39941.67
L7a
7514.98
14861.86
22208.75
L9
5180.13
19806.03
34431.94
L10
3369.67
21600.01
39830.36
L13
3243.83
20220.69
37197.54
L14
4658.33
18834.85
33111.36
L17
4636.41
20401.27
36166.13
L18
3398.17
21568.12
39738.07
L21
3229.88
20921.88
38613.89
L22
3378.08
21079.69
38781.29
L25
5006.87
20319.15
35631.42
L27
3266.91
21214.63
39162.34
L30
3691.32
21688.74
39686.16
L31
3847.59
21868.38
39889.16
L34
3285.78
20254.42
37223.06
L35
4149.15
16726.55
29303.95
S2
5134.01
22557.35
39980.68
S3
5850.54
22828.03
39805.53
S6
3998.49
21941.34
39884.2
S7
3586.02
21729.97
39873.93
S10
3614.24
21765.9
39917.57
S11
2613.1
20994.68
39376.26
S14
3922.51
20206.28
36490.05
S15
3340.66
20018.57
36696.49
S18
4317.2
20369.97
36422.73
S19
2970.35
16855.88
30741.4
S22
4805.56
6911.09
9016.62
S31e
4895.28
4986.39
5077.49

APÊNDICE 2 – Relação do número de dados faltantes na base de dados presumida

Proteínas Ribossomais	**L1**	**L2**	**L3**	**L4**	**L5**	**L6**	**L7**	**L7a**
40k-mais								
Total	8744	8664	3834	4024	6662	8065	29508	28429
*	27%	27%	12%	13%	21%	25%	92%	89%
40k-menos								
Total	8775	8754	3905	4062	6666	8067	29508	28429
*	27%	27%	12%	13%	21%	25%	92%	89%
Diferença								
Total	31	90	71	38	4	2	0	0
*	0%	0%	0%	0%	0%	0%	0%	0%
**	0,36%	1,04%	1,86%	0,95%	0,07%	0,03%	0%	0%
*diferença em relação à base total de 32.026 registros								
**porcentagem do aumento de dados faltantes de 40k-mais para 40k-menos Obs: nestes resultados também estavam presentes Filos e Famílias, totalizando 32.026 registros								

L7ae	**L7/L12**	**L9**	**L10**	**L11**	**L12**	**L13**	**L14**	**L15**	**L16**
27231	10408	9338	7402	8883	31327	8771	9623	8497	9135
85%	32%	29%	23%	28%	98%	27%	30%	27%	29%
27232	10409	9365	7404	8894	31337	8775	9623	8511	9323
85%	33%	29%	23%	28%	98%	27%	30%	27%	29%
1	1	27	2	11	10	4	0	14	188
0%	1%	0%	0%	0%	0%	0%	0%	0%	0%
0,01%	0,01%	0,29%	0,03%	0,13%	0,04%	0,05%	0%	0,17%	2,06%

L17	L18	L19	L20	L21	L22	L23	L24	L25	L27
9173	8808	8777	9711	8935	9061	6984	8813	11400	9599
29%	28%	27%	30%	28%	28%	22%	28%	36%	30%
9180	8811	8784	9716	8949	9061	6987	8819	11400	9603
29%	28%	27%	30%	28%	28%	22%	28%	36%	30%
7	3	7	5	14	0	3	6	0	4
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
0,08%	0,04%	0,08%	0,06%	0,16%	0%	0,05%	0,07%	0%	0,05%

L28	L29	L30	L31	L32	L33	L34	L35	L36	S1
10700	10166	11773	8937	11035	11264	14841	10852	15049	10823
33%	32%	37%	28%	34%	35%	46%	34%	47%	34%
10704	10170	11774	8948	11035	11266	14841	10854	15053	30878
33%	32%	37%	28%	34%	35%	46%	34%	47%	96%
4	4	1	11	0	2	0	2	4	20055
0%	0%	0%	0%	0%	0%	0%	0%	0%	62%
0,04%	0,04%	0,01%	0,13%	0%	0,02%	0%	0,02%	0,03%	185,3%

S2	S3	S4	S5	S6	S7	S8	S9	S10	S11
8414	8762	8620	9124	9029	9226	6739	6885	10536	6072
26%	27%	27%	28%	28%	29%	21%	21%	33%	19%
8495	8770	8624	9159	9069	9243	6739	6913	10553	6073
27%	27%	27%	29%	28%	29%	21%	22%	33%	19%
81	8	4	35	40	17	0	28	17	1
1%	0%	0%	1%	0%	0%	0%	1%	0%	0%
0,97%	0,1%	0,05%	0,39%	0,45%	0,19%	0%	0,41%	0,17%	0,02%

S12	S13	S14	S15	S16	S17	S18	S19	S20	S21
9897	4806	9653	9056	9479	4584	10485	9340	9842	15139
31%	15%	30%	28%	30%	14%	33%	29%	31%	47%
9933	4807	9656	9066	9482	4592	10490	9340	9843	15139
31%	15%	30%	28%	30%	14%	33%	29%	31%	47%
36	1	3	10	3	8	5	0	1	0
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
0,37%	0,03%	0,04%	0,12%	0,04%	0,18%	0,05%	0%	0,02%	0%

S22	S31e
31481	32021
98%	100%
31481	32021
98%	100%
0	0
0%	0%
0%	0%

APÊNDICE 3 – Matriz de correlação das proteínas ribossomais

Correlation Matrix

	L1	L2	L3	L4	L5	L6	L7	L7a	L7ae	L7/L12	L9	L10
L1	1	0,05	0,04	0,18	0,24	0,18	0,05	-0,01	0,08	0,19	0,04	0,05
L2	0,05	1	-0,26	0,23	0,11	-0,02	0	0,01	-0,11	0,03	0,05	-0,26
L3	0,04	-0,26	1	-0,13	0,12	0,22	0,14	-0,02	0,23	0,05	0,29	0,42
L4	0,18	0,23	-0,13	1	0,23	0,09	0,05	0	-0,07	0,14	0,04	-0,16
L5	0,24	0,11	0,12	0,23	1	0,29	0,21	0	0,23	0,18	0,23	0,08
L6	0,18	-0,02	0,22	0,09	0,29	1	0,11	-0,04	0,14	0,15	0,17	0,2
L7	0,05	0	0,14	0,05	0,21	0,11	1	0,02	0,28	0,05	0,17	0,01
L7a	-0,01	0,01	-0,02	0	0	-0,04	0,02	1	0,01	-0,02	0,01	0
L7ae	0,08	-0,11	0,23	-0,07	0,23	0,14	0,28	0,01	1	0,05	0,06	0,17
L7/L12	0,19	0,03	0,05	0,14	0,18	0,15	0,05	-0,02	0,05	1	0,07	0,11
L9	0,04	0,05	0,29	0,04	0,23	0,17	0,17	0,01	0,06	0,07	1	0,06
L10	0,05	-0,26	0,42	-0,16	0,08	0,2	0,01	0	0,17	0,11	0,06	1
L11	0,04	-0,03	0,12	0	0,06	0,08	0,02	-0,01	0,02	0,04	0,18	0,11
L12	0,27	0,13	-0,01	0,19	0,2	0,14	0,02	0	-0,02	0,08	0,11	0
L13	0,2	0,03	0,2	0,11	0,27	0,24	0,12	0	0,12	0,09	0,31	0,06
L14	0,04	-0,03	0,13	0	0,09	0,04	0,09	0,01	0,15	0,03	0,02	0,07
L15	0,14	-0,14	0,38	-0,04	0,19	0,26	0,03	0	0,11	0,12	0,27	0,3
L16	0	-0,09	0,28	-0,15	0	0,23	-0,03	0	0,08	0,01	0	0,31
L17	0,12	0,05	0,05	0,19	0,28	0,06	0,06	0,03	0,08	0,19	0,05	0,17
L18	0,13	-0,07	0,29	0,04	0,2	0,25	0,12	-0,02	0,21	0,15	0,03	0,21
L19	0,15	-0,12	0,34	-0,02	0,21	0,25	0,1	0	0,19	0,21	0,42	0,28
L20	0,04	0,07	0,06	0,07	0,12	-0,01	-0,03	0	-0,01	0,05	0,09	0,05
L21	0,1	0,05	0,16	0,1	0,19	0,13	0,2	0	0,05	0,12	0,45	0,02
L22	0,12	-0,01	0,19	0,1	0,23	0,17	0,08	-0,01	0,13	0,25	0,08	0,23
L23	0,12	0,16	-0,02	0,18	0,08	0,02	0,02	0,03	0,04	0,18	0	-0,05
L24	0,18	0,02	0,21	0,18	0,23	0,22	0,11	-0,02	0,13	0,14	0,03	0,14
L25	0,04	0,08	0,14	0,11	0,14	0,27	0	-0,03	-0,05	0,18	0,22	0,08
L27	0,17	-0,01	0,16	0,11	0,09	0,27	0,02	-0,01	0,06	0,08	0,08	0,03
L28	0,11	-0,02	0,15	0,08	0,26	0	0,15	0	0,11	0,11	0,4	0,08
L29	0,17	0,08	0,15	0,14	0,25	0,22	-0,01	-0,04	-0,01	0,18	0,07	0,09
L30	-0,03	-0,25	0,35	-0,19	-0,01	0,07	0,03	-0,01	0,17	0,03	0,03	0,34
L31	0,03	-0,09	0,09	-0,02	0,03	0,11	0,04	0,02	0,11	-0,01	-0,04	0,05
L32	-0,08	-0,4	0,56	-0,29	-0,07	0,11	0,04	0,02	0,21	0,04	0,05	0,59
L33	0,04	0,01	0,01	0,02	0,05	-0,02	0,06	0	0,09	0,05	0,09	0,02
L34	0,17	-0,1	0,21	0,04	0,24	0,28	0,22	0,01	0,18	0,08	0,15	0,2
L35	0,23	-0,07	0,17	0,02	0,22	0,27	0,04	0	0,13	0,12	0,16	0,09
L36	0,1	-0,04	0,11	0,07	0,14	0	0,17	0,01	0,21	0,02	0,09	0,03
S1	0,07	0,04	-0,02	0,01	0,04	0,05	0	0,01	-0,01	0,08	0,01	0,03
S2	0,08	0,21	-0,04	0,23	0,22	0,08	0,03	0,02	-0,04	0,17	0,11	0,01
S3	0,01	0,09	0,02	0,13	0,17	-0,1	0,05	0,01	0,02	0,09	-0,04	0,13
S4	0,03	0,03	0,07	-0,01	0,06	-0,02	0,04	0,03	0,05	0,03	0,06	0,1
S5	0,11	-0,02	0,36	0,12	0,38	0,13	0,13	-0,01	0,13	0,2	0,21	0,29
S6	0,02	-0,07	0,12	-0,03	0,09	-0,03	0,1	0	0,12	0,03	0,19	0,07
S7	-0,02	-0,22	0,41	-0,17	-0,04	0,16	0,04	0	0,19	0,02	0,05	0,37
S8	0,07	0,09	0,01	0,11	0,1	0,23	0,03	0,03	0,07	0,06	0,08	-0,02
S9	0,17	0,09	0,16	0,17	0,36	0,06	0,11	-0,03	0,05	0,28	0,3	0,1
S10	0,23	-0,02	0,14	0,13	0,19	0,15	0,05	0,01	0,08	0,18	0,07	0,08
S11	0,1	0,08	0,05	0,1	0,17	0,09	0,02	-0,01	0,04	0,14	0,03	0,06
S12	0,05	-0,07	0,14	-0,04	-0,05	0,24	-0,02	-0,02	0,05	0,02	0	0,06
S13	0,11	-0,27	0,52	-0,09	0,19	0,34	0,09	-0,01	0,22	0,19	0,11	0,49
S14	-0,03	0,09	-0,09	0,03	0,04	-0,15	0,09	-0,05	0,01	-0,05	0,09	-0,2
S15	0,04	-0,25	0,45	-0,18	0,03	0,19	0,08	0	0,2	0,12	0,05	0,38
S16	0,03	0,09	0,07	0,16	0,27	0,11	0,06	0	0,03	0,2	0,2	0,13
S17	0,08	0,06	0,07	0,15	0,22	0,1	0,02	0,02	0,09	0,17	-0,09	0,09
S18	0,08	0	0,09	0,06	0,09	0,17	0,03	0,01	0,11	0,1	0,11	0,1
S19	0	-0,31	0,45	-0,21	0,05	0,16	0,08	0	0,23	0,09	0	0,42
S20	0,21	-0,01	0,09	0,15	0,26	0,21	0,02	-0,01	0,02	0,21	0,2	0,15
S21	-0,04	-0,04	0,08	0	0,12	-0,17	0,16	0	0,09	0,01	0,35	0,02
S22	0	0	-0,02	-0,01	0	0	0	0	0	-0,02	0	0
S31e	0	0	0	0	0	0	0	0	0	0	0	0

L11	L12	L13	L14	L15	L16	L17	L18	L19	L20	L21	L22	L23
0,04	0,27	0,2	0,04	0,14	0	0,12	0,13	0,15	0,04	0,1	0,12	0,12
-0,03	0,13	0,03	-0,03	-0,14	-0,09	0,05	-0,07	-0,12	0,07	0,05	-0,01	0,16
0,12	-0,01	0,2	0,13	0,38	0,28	0,05	0,29	0,34	0,06	0,16	0,19	-0,02
0	0,19	0,11	0	-0,04	-0,15	0,19	0,04	-0,02	0,07	0,1	0,1	0,18
0,06	0,2	0,27	0,09	0,19	0	0,28	0,2	0,21	0,12	0,19	0,23	0,08
0,08	0,14	0,24	0,04	0,26	0,23	0,06	0,25	0,25	-0,01	0,13	0,17	0,02
0,02	0,02	0,12	0,09	0,03	-0,03	0,06	0,12	0,1	-0,03	0,2	0,08	0,02
-0,01	0	0	0,01	0	0	0,03	-0,02	0	0	0	-0,01	0,03
0,02	-0,02	0,12	0,15	0,11	0,08	0,08	0,21	0,19	-0,01	0,05	0,13	0,04
0,04	0,08	0,09	0,03	0,12	0,01	0,19	0,15	0,21	0,05	0,12	0,25	0,18
0,18	0,11	0,31	0,02	0,27	0	0,05	0,03	0,42	0,09	0,45	0,08	0
0,11	0	0,06	0,07	0,3	0,31	0,17	0,21	0,28	0,05	0,02	0,23	-0,05
1	0,04	0,11	0,04	0,14	0,04	0,06	0,07	0,18	0,05	0,15	0,04	0,04
0,04	1	0,21	0,05	0,09	-0,02	0,11	0,02	0,08	0,03	0,17	0,08	0,09
0,11	0,21	1	0,08	0,26	0,09	0,17	0,19	0,27	0,1	0,26	0,07	0,08
0,04	0,05	0,08	1	0,08	-0,01	0,11	0,09	0,12	0,01	0,06	0,09	0,04
0,14	0,09	0,26	0,08	1	0,16	0,13	0,23	0,38	0,08	0,2	0,14	0,03
0,04	-0,02	0,09	-0,01	0,16	1	-0,03	0,18	0,1	0	-0,04	0,07	-0,06
0,06	0,11	0,17	0,11	0,13	-0,03	1	0,12	0,04	0,19	0,08	0,45	0,05
0,07	0,02	0,19	0,09	0,23	0,18	0,12	1	0,24	0,1	0,07	0,16	0,07
0,18	0,08	0,27	0,12	0,38	0,1	0,04	0,24	1	0,09	0,37	0,17	0,07
0,05	0,03	0,1	0,01	0,08	0	0,19	0,1	0,09	1	0,02	0,18	-0,02
0,15	0,17	0,26	0,06	0,2	-0,04	0,08	0,07	0,37	0,02	1	0,05	0,06
0,04	0,08	0,07	0,09	0,14	0,07	0,45	0,16	0,17	0,18	0,05	1	0,01
0,04	0,09	0,08	0,04	0,03	-0,06	0,05	0,07	0,07	-0,02	0,06	0,01	1
0,07	0,3	0,28	0,19	0,23	0,1	0,21	0,23	0,21	0,04	0,12	0,11	0,21
0,03	0,02	0,1	-0,13	0,14	0,19	0,13	0,08	0,14	0,09	0,07	0,21	-0,07
0,05	0,25	0,21	0,06	0,18	0,15	-0,05	0,16	0,14	0,02	0,1	-0,03	0,05
0,11	0,28	0,26	0,17	0,19	-0,19	0,26	0,01	0,33	0,06	0,36	0,23	0,11
0,06	0,17	0,18	0	0,21	0,06	0,25	0,17	0,15	0,18	0,13	0,23	0,13
0,06	-0,08	0,02	0,14	0,22	0,16	0,05	0,17	0,22	0,01	0	0,15	-0,07
-0,02	0,03	0,08	0,04	0,1	0,03	-0,03	0,09	0,07	-0,05	0,02	-0,01	0,04
0,09	-0,18	0,01	0,14	0,34	0,28	0,04	0,22	0,31	-0,01	0	0,17	-0,09
0	0,06	0,04	0,05	0,01	-0,02	0,13	0,03	0,07	0,02	0,08	0,11	0,06
0,09	0,16	0,28	0,07	0,3	0,08	0,11	0,26	0,32	0,02	0,25	0,17	0,07
0,08	0,24	0,25	0	0,25	0,1	-0,03	0,19	0,29	0	0,19	0	0,07
0,02	0,18	0,16	0,21	0,07	-0,08	0,07	0,1	0,15	-0,01	0,13	-0,02	0,11
0,02	0,01	0,03	0	0,02	0,02	0,02	0	0,04	-0,02	0,07	0,02	-0,01
0,02	0,03	0,1	-0,05	0,06	-0,01	0,36	0,03	-0,01	0,19	0,03	0,25	0,09
0,02	0,01	0,01	0,1	0,01	-0,09	0,47	0,04	0,03	0,2	0,03	0,35	0,07
0,02	-0,02	0,03	0,09	0,05	-0,01	0,06	0,05	0,06	0,07	0,02	0,07	0,03
0,09	0,04	0,18	0,11	0,31	0,07	0,5	0,23	0,28	0,27	0,13	0,52	0,04
0,09	0,05	0,09	0,12	0,11	-0,12	-0,05	0	0,28	-0,04	0,25	-0,05	0,21
0,06	-0,07	0,07	0,14	0,22	0,26	0,02	0,2	0,21	-0,02	0,01	0,11	-0,08
0,01	0,03	0,15	0,03	0,08	0,15	0,03	0,1	0,07	0,03	0,02	0,04	0,02
0,08	0,11	0,19	0,06	0,22	-0,06	0,41	0,18	0,25	0,22	0,24	0,3	0,18
0,05	0,21	0,15	0,1	0,16	0,02	0,06	0,17	0,23	0	0,16	0,05	0,17
0,02	0,12	0,05	0,07	0,05	0,04	0,2	0,07	0,05	0,07	0,04	0,29	0,06
0,02	-0,01	0,05	-0,07	0,11	0,23	-0,14	0,12	0,08	-0,03	0,02	-0,01	-0,03
0,12	-0,05	0,14	0,15	0,37	0,26	0,19	0,3	0,36	0,07	0,07	0,38	-0,05
-0,02	0,07	0,05	0,05	-0,08	-0,2	-0,04	-0,14	-0,01	-0,03	0,15	-0,15	0,07
0,07	-0,06	0,11	0,12	0,26	0,25	0,07	0,28	0,26	0,01	0,05	0,18	0,01
0,06	0,02	0,14	0,03	0,14	0,02	0,63	0,07	0,11	0,21	0,2	0,51	0,05
-0,01	0,05	0,11	0,08	0,07	0,05	0,48	0,16	-0,04	0,2	-0,08	0,56	-0,01
0,08	0,03	0,13	0	0,11	0,05	0,16	0,1	0,19	0,04	0,11	0,18	0,11
0,05	-0,14	0,04	0,11	0,24	0,24	0,02	0,26	0,22	0,03	-0,04	0,19	-0,11
0,14	0,23	0,21	0,11	0,24	-0,05	0,12	0,12	0,31	0,01	0,27	0,08	0,13
0,07	0,01	0,08	0,16	0,04	-0,24	0,12	-0,08	0,27	0,08	0,31	0,06	0,03
0	0	0	0	-0,01	0	0	0	0	0	0	0	0
0	0	0	0	0,01	0	0	-0,03	0	-0,03	0	0	0

L24	L25	L27	L28	L29	L30	L31	L32	L33	L34	L35	L36	S1
0,18	0,04	0,17	0,11	0,17	-0,03	0,03	-0,08	0,04	0,17	0,23	0,1	0,07
0,02	0,08	-0,01	-0,02	0,08	-0,25	-0,09	-0,4	0,01	-0,1	-0,07	-0,04	0,04
0,21	0,14	0,16	0,15	0,15	0,35	0,09	0,56	0,01	0,21	0,17	0,11	-0,02
0,18	0,11	0,11	0,08	0,14	-0,19	-0,02	-0,29	0,02	0,04	0,02	0,07	0,01
0,23	0,14	0,09	0,26	0,25	-0,01	0,03	-0,07	0,05	0,24	0,22	0,14	0,04
0,22	0,27	0,27	0	0,22	0,07	0,11	0,11	-0,02	0,28	0,27	0	0,05
0,11	0	0,02	0,15	-0,01	0,03	0,04	0,04	0,06	0,22	0,04	0,17	0
-0,02	-0,03	-0,01	0	-0,04	-0,01	0,02	0,02	0	0,01	0	0,01	0,01
0,13	-0,05	0,06	0,11	-0,01	0,17	0,11	0,21	0,09	0,18	0,13	0,21	-0,01
0,14	0,18	0,08	0,11	0,18	0,03	-0,01	0,04	0,05	0,08	0,12	0,02	0,08
0,03	0,22	0,08	0,4	0,07	0,03	-0,04	0,05	0,09	0,15	0,16	0,09	0,01
0,14	0,08	0,03	0,08	0,09	0,34	0,05	0,59	0,02	0,2	0,09	0,03	0,03
0,07	0,03	0,05	0,11	0,06	0,06	-0,02	0,09	0	0,09	0,08	0,02	0,02
0,3	0,02	0,25	0,28	0,17	-0,08	0,03	-0,18	0,06	0,16	0,24	0,18	0,01
0,28	0,1	0,21	0,26	0,18	0,02	0,08	0,01	0,04	0,28	0,25	0,16	0,03
0,19	-0,13	0,06	0,17	0	0,14	0,04	0,14	0,05	0,07	0	0,21	0
0,23	0,14	0,18	0,19	0,21	0,22	0,1	0,34	0,01	0,3	0,25	0,07	0,02
0,1	0,19	0,15	-0,19	0,06	0,16	0,03	0,28	-0,02	0,08	0,1	-0,08	0,02
0,21	0,13	-0,05	0,26	0,25	0,05	-0,03	0,04	0,13	0,11	-0,03	0,07	0,02
0,23	0,08	0,16	0,01	0,17	0,17	0,09	0,22	0,03	0,26	0,19	0,1	0
0,21	0,14	0,14	0,33	0,15	0,22	0,07	0,31	0,07	0,32	0,29	0,15	0,04
0,04	0,09	0,02	0,06	0,18	0,01	-0,05	-0,01	0,02	0,02	0	-0,01	-0,02
0,12	0,07	0,1	0,36	0,13	0	0,02	0	0,08	0,25	0,19	0,13	0,07
0,11	0,21	-0,03	0,23	0,23	0,15	-0,01	0,17	0,11	0,17	0	-0,02	0,02
0,21	-0,07	0,05	0,11	0,13	-0,07	0,04	-0,09	0,06	0,07	0,07	0,11	-0,01
1	0,09	0,33	0,18	0,22	0,09	0,11	0,11	-0,08	0,21	0,19	0,29	0,05
0,09	1	0,15	-0,14	0,17	-0,03	-0,06	0,03	-0,17	0,02	0,1	-0,25	0,01
0,33	0,15	1	-0,08	0,13	0,01	0,09	0,04	-0,02	0,16	0,26	0,25	-0,01
0,18	-0,14	-0,08	1	0,06	0,09	0,01	0,05	0,17	0,21	0,1	0,26	0,06
0,22	0,17	0,13	0,06	1	0,02	0,07	0,01	0,03	0,18	0,2	-0,05	0,03
0,09	-0,03	0,01	0,09	0,02	1	0,04	0,46	0,06	0,11	0,05	0,06	0
0,11	-0,06	0,09	0,01	0,07	0,04	1	0,09	0,06	0,18	0,15	0,09	-0,01
0,11	0,03	0,04	0,05	0,01	0,46	0,09	1	0,02	0,15	0,07	0,05	-0,02
-0,08	-0,17	-0,02	0,17	0,03	0,06	0,06	0,02	1	0,07	0,03	0,06	0,04
0,21	0,02	0,16	0,21	0,18	0,11	0,18	0,15	0,07	1	0,35	0,19	0
0,19	0,1	0,26	0,1	0,2	0,05	0,15	0,07	0,03	0,35	1	0,08	0,02
0,29	-0,25	0,25	0,26	-0,05	0,06	0,09	0,05	0,06	0,19	0,08	1	0,01
0,05	0,01	-0,01	0,06	0,03	0	-0,01	-0,02	0,04	0	0,02	0,01	1
0,08	0,33	-0,04	0,01	0,26	-0,08	-0,06	-0,07	0,02	-0,05	-0,08	-0,11	0,01
0,04	-0,01	-0,23	0,25	0,15	0,06	-0,09	0,05	0,1	0	-0,16	0,01	0,02
0	-0,06	-0,08	0,09	0,03	0,05	-0,02	0,05	0,01	0,03	-0,03	0,03	-0,01
0,19	0,2	-0,05	0,3	0,39	0,2	-0,01	0,25	0,09	0,23	0,07	0,02	0
0,03	-0,2	-0,06	0,37	0,01	0,09	0,08	0,07	0,08	0,13	0,07	0,2	-0,01
0,18	0,03	0,19	0	0	0,3	0,08	0,49	0,02	0,15	0,08	0,12	0
0,07	0,24	0,14	-0,13	0,07	-0,07	0,05	-0,06	-0,06	0,07	0,11	0	0,01
0,17	0,19	-0,03	0,32	0,29	0,03	-0,06	0,01	0,05	0,13	0,1	0,07	0,02
0,26	-0,02	0,27	0,18	0,12	0,06	0,02	0,04	0,04	0,13	0,21	0,17	0,11
0,08	0,09	0	0,12	0,11	0,03	-0,01	-0,02	0,04	0,08	0,06	0	0,03
0,06	0,24	0,26	-0,27	0,04	0,03	0,11	0,12	-0,06	0,1	0,13	-0,05	0,01
0,21	0,25	0,1	0,1	0,21	0,37	0,1	0,58	0	0,24	0,17	0,01	0,04
0,07	-0,18	-0,11	0,37	-0,1	-0,09	0,08	-0,19	0,06	-0,03	0,01	0,15	0
0,18	0,04	0,16	0,03	0,06	0,34	0,1	0,5	0,05	0,19	0,13	0,15	-0,04
0,12	0,29	-0,09	0,26	0,23	0,02	-0,03	0,05	0,12	0,1	-0,08	-0,05	-0,01
0,12	0,2	-0,04	0,09	0,23	0,06	-0,02	0,03	0,02	0,08	-0,01	-0,07	-0,02
0,06	0,14	0,07	0,05	0,08	0,02	0,02	0,06	0,05	0,15	0,1	0,07	-0,01
0,09	0,23	0,04	0	0,03	0,32	0,08	0,5	-0,04	0,18	0,08	0,04	-0,01
0,16	-0,01	0,06	0,35	0,19	0,07	0,07	0,03	0,06	0,28	0,29	0,08	0,14
-0,05	-0,21	-0,21	0,55	-0,12	0,08	-0,09	0,03	0,1	0,08	-0,08	0,22	0,01
0	0	0	-0,01	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	-0,01

S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14
0,08	0,01	0,03	0,11	0,02	-0,02	0,07	0,17	0,23	0,1	0,05	0,11	-0,03
0,21	0,09	0,03	-0,02	-0,07	-0,22	0,09	0,09	-0,02	0,08	-0,07	-0,27	0,09
-0,04	0,02	0,07	0,36	0,12	0,41	0,01	0,16	0,14	0,05	0,14	0,52	-0,09
0,23	0,13	-0,01	0,12	-0,03	-0,17	0,11	0,17	0,13	0,1	-0,04	-0,09	0,03
0,22	0,17	0,06	0,38	0,09	-0,04	0,1	0,36	0,19	0,17	-0,05	0,19	0,04
0,08	-0,1	-0,02	0,13	-0,03	0,16	0,23	0,06	0,15	0,09	0,24	0,34	-0,15
0,03	0,05	0,04	0,13	0,1	0,04	0,03	0,11	0,05	0,02	-0,02	0,09	0,09
0,02	0,01	0,03	-0,01	0	0	0,03	-0,03	0,01	-0,01	-0,02	-0,01	-0,05
-0,04	0,02	0,05	0,13	0,12	0,19	0,07	0,05	0,08	0,04	0,05	0,22	0,01
0,17	0,09	0,03	0,2	0,03	0,02	0,06	0,28	0,18	0,14	0,02	0,19	-0,05
0,11	-0,04	0,06	0,21	0,19	0,05	0,08	0,3	0,07	0,03	0	0,11	0,09
0,01	0,13	0,1	0,29	0,07	0,37	-0,02	0,1	0,08	0,06	0,06	0,49	-0,2
0,02	0,02	0,02	0,09	0,09	0,06	0,01	0,08	0,05	0,02	0,02	0,12	-0,02
0,03	0,01	-0,02	0,04	0,05	-0,07	0,03	0,11	0,21	0,12	-0,01	-0,05	0,07
0,1	0,01	0,03	0,18	0,09	0,07	0,15	0,19	0,15	0,05	0,05	0,14	0,05
-0,05	0,1	0,09	0,11	0,12	0,14	0,03	0,06	0,1	0,07	-0,07	0,15	0,05
0,06	0,01	0,05	0,31	0,11	0,22	0,08	0,22	0,16	0,05	0,11	0,37	-0,08
-0,01	-0,09	-0,01	0,07	-0,12	0,26	0,15	-0,06	0,02	0,04	0,23	0,26	-0,2
0,36	0,47	0,06	0,5	-0,05	0,02	0,03	0,41	0,06	0,2	-0,14	0,19	-0,04
0,03	0,04	0,05	0,23	0	0,2	0,1	0,18	0,17	0,07	0,12	0,3	-0,14
-0,01	0,03	0,06	0,28	0,28	0,21	0,07	0,25	0,23	0,05	0,08	0,36	-0,01
0,19	0,2	0,07	0,27	-0,04	-0,02	0,03	0,22	0	0,07	-0,03	0,07	-0,03
0,03	0,03	0,02	0,13	0,25	0,01	0,02	0,24	0,16	0,04	0,02	0,07	0,15
0,25	0,35	0,07	0,52	-0,05	0,11	0,04	0,3	0,05	0,29	-0,01	0,38	-0,15
0,09	0,07	0,03	0,04	0,21	-0,08	0,02	0,18	0,17	0,06	-0,03	-0,05	0,07
0,08	0,04	0	0,19	0,03	0,18	0,07	0,17	0,26	0,08	0,06	0,21	0,07
0,33	-0,01	-0,06	0,2	-0,2	0,03	0,24	0,19	-0,02	0,09	0,24	0,25	-0,18
-0,04	-0,23	-0,08	-0,05	-0,06	0,19	0,14	-0,03	0,27	0	0,26	0,1	-0,11
0,01	0,25	0,09	0,3	0,37	0	-0,13	0,32	0,18	0,12	-0,27	0,1	0,37
0,26	0,15	0,03	0,39	0,01	0	0,07	0,29	0,12	0,11	0,04	0,21	-0,1
-0,08	0,06	0,05	0,2	0,09	0,3	-0,07	0,03	0,06	0,03	0,03	0,37	-0,09
-0,06	-0,09	-0,02	-0,01	0,08	0,08	0,05	-0,06	0,02	-0,01	0,11	0,1	0,08
-0,07	0,05	0,05	0,25	0,07	0,49	-0,06	0,01	0,04	-0,02	0,12	0,58	-0,19
0,02	0,1	0,01	0,09	0,08	0,02	-0,06	0,05	0,04	0,04	-0,06	0	0,06
-0,05	0	0,03	0,23	0,13	0,15	0,07	0,13	0,13	0,08	0,1	0,24	-0,03
-0,08	-0,16	-0,03	0,07	0,07	0,08	0,11	0,1	0,21	0,06	0,13	0,17	0,01
-0,11	0,01	0,03	0,02	0,2	0,12	0	0,07	0,17	0	-0,05	0,01	0,15
0,01	0,02	-0,01	0	-0,01	0	0,01	0,02	0,11	0,03	0,01	0,04	0
1	0,28	0,06	0,28	-0,09	-0,09	0,12	0,29	-0,04	0,09	-0,03	0,07	-0,08
0,28	1	0,11	0,45	0,05	-0,04	-0,11	0,31	-0,01	0,18	-0,27	0,13	0,03
0,06	0,11	1	0,11	0,08	0,03	0,01	0,09	0,01	0,03	-0,09	0,04	0
0,28	0,45	0,11	1	0,03	0,13	-0,03	0,54	0,1	0,22	-0,11	0,43	-0,08
-0,09	0,05	0,08	0,03	1	0,04	-0,1	0,1	0,15	-0,02	-0,15	0,03	0,21
-0,09	-0,04	0,03	0,13	0,04	1	0,02	-0,02	0,1	0	0,16	0,37	-0,14
0,12	-0,11	0,01	-0,03	-0,1	0,02	1	0,01	0,01	0,03	0,17	0,07	-0,13
0,29	0,31	0,09	0,54	0,1	-0,02	0,01	1	0,14	0,23	-0,14	0,2	0,09
-0,04	-0,01	0,01	0,1	0,15	0,1	0,01	0,14	1	0,07	0,08	0,12	0,04
0,09	0,18	0,03	0,22	-0,02	0	0,03	0,23	0,07	1	-0,05	0,16	-0,02
-0,03	-0,27	-0,09	-0,11	-0,15	0,16	0,17	-0,14	0,08	-0,05	1	0,14	-0,2
0,07	0,13	0,04	0,43	0,03	0,37	0,07	0,2	0,12	0,16	0,14	1	-0,25
-0,08	0,03	0	-0,08	0,21	-0,14	-0,13	0,09	0,04	-0,02	-0,2	-0,25	1
-0,08	-0,03	0,05	0,18	0,08	0,34	0	0,09	0,13	0,09	0,14	0,44	-0,12
0,4	0,47	0,02	0,55	-0,05	0,01	0,02	0,48	-0,01	0,22	-0,09	0,21	-0,01
0,25	0,33	0,05	0,44	-0,18	0,01	0,15	0,21	0,01	0,27	-0,07	0,26	-0,09
0,09	0,11	0,01	0,05	0,06	0,08	0,09	0,03	0,04	0,1	0,06	0,15	-0,1
-0,08	0	0,06	0,23	0,02	0,35	0,07	0,06	0,04	0,06	0,13	0,52	-0,18
0	0,11	0,06	0,18	0,23	0,01	-0,02	0,16	0,23	0,11	-0,12	0,18	0,01
-0,02	0,23	0,12	0,15	0,41	-0,02	-0,17	0,25	0,08	0,03	-0,31	-0,03	0,28
-0,01	0	-0,05	0	0	0	0	0	0	0	0	0	0
0	0	-0,02	0	0	0	0	0	0	0	0	0	0

S15	S16	S17	S18	S19	S20	S21	S22	S31e
0,04	0,03	0,08	0,08	0	0,21	-0,04	0	0
-0,25	0,09	0,06	0	-0,31	-0,01	-0,04	0	0
0,45	0,07	0,07	0,09	0,45	0,09	0,08	-0,02	0
-0,18	0,16	0,15	0,06	-0,21	0,15	0	-0,01	0
0,03	0,27	0,22	0,09	0,05	0,26	0,12	0	0
0,19	0,11	0,1	0,17	0,16	0,21	-0,17	0	0
0,08	0,06	0,02	0,03	0,08	0,02	0,16	0	0
0	0	0,02	0,01	0	-0,01	0	0	0
0,2	0,03	0,09	0,11	0,23	0,02	0,09	0	0
0,12	0,2	0,17	0,1	0,09	0,21	0,01	-0,02	0
0,05	0,2	-0,09	0,11	0	0,2	0,35	0	0
0,38	0,13	0,09	0,1	0,42	0,15	0,02	0	0
0,07	0,06	-0,01	0,08	0,05	0,14	0,07	0	0
-0,06	0,02	0,05	0,03	-0,14	0,23	0,01	0	0
0,11	0,14	0,11	0,13	0,04	0,21	0,08	0	0
0,12	0,03	0,08	0	0,11	0,11	0,16	0	0
0,26	0,14	0,07	0,11	0,24	0,24	0,04	-0,01	0,01
0,25	0,02	0,05	0,05	0,24	-0,05	-0,24	0	0
0,07	0,63	0,48	0,16	0,02	0,12	0,12	0	0
0,28	0,07	0,16	0,1	0,26	0,12	-0,08	0	-0,03
0,26	0,11	-0,04	0,19	0,22	0,31	0,27	0	0
0,01	0,21	0,2	0,04	0,03	0,01	0,08	0	-0,03
0,05	0,2	-0,08	0,11	-0,04	0,27	0,31	0	0
0,18	0,51	0,56	0,18	0,19	0,08	0,06	0	0
0,01	0,05	-0,01	0,11	-0,11	0,13	0,03	0	0
0,18	0,12	0,12	0,06	0,09	0,16	-0,05	0	0
0,04	0,29	0,2	0,14	0,23	-0,01	-0,21	0	0
0,16	-0,09	-0,04	0,07	0,04	0,06	-0,21	0	0
0,03	0,26	0,09	0,05	0	0,35	0,55	-0,01	0
0,06	0,23	0,23	0,08	0,03	0,19	-0,12	0	0
0,34	0,02	0,06	0,02	0,32	0,07	0,08	0	0
0,1	-0,03	-0,02	0,02	0,08	0,07	-0,09	0	0
0,5	0,05	0,03	0,06	0,5	0,03	0,03	0	0
0,05	0,12	0,02	0,05	-0,04	0,06	0,1	0	0
0,19	0,1	0,08	0,15	0,18	0,28	0,08	0	0
0,13	-0,08	-0,01	0,1	0,08	0,29	-0,08	0	0
0,15	-0,05	-0,07	0,07	0,04	0,08	0,22	0	0
-0,04	-0,01	-0,02	-0,01	-0,01	0,14	0,01	0	-0,01
-0,08	0,4	0,25	0,09	-0,08	0	-0,02	-0,01	0
-0,03	0,47	0,33	0,11	0	0,11	0,23	0	0
0,05	0,02	0,05	0,01	0,06	0,06	0,12	-0,05	-0,02
0,18	0,55	0,44	0,05	0,23	0,18	0,15	0	0
0,08	-0,05	-0,18	0,06	0,02	0,23	0,41	0	0
0,34	0,01	0,01	0,08	0,35	0,01	-0,02	0	0
0	0,02	0,15	0,09	0,07	-0,02	-0,17	0	0
0,09	0,48	0,21	0,03	0,06	0,16	0,25	0	0
0,13	-0,01	0,01	0,04	0,04	0,23	0,08	0	0
0,09	0,22	0,27	0,1	0,06	0,11	0,03	0	0
0,14	-0,09	-0,07	0,06	0,13	-0,12	-0,31	0	0
0,44	0,21	0,26	0,15	0,52	0,18	-0,03	0	0
-0,12	-0,01	-0,09	-0,1	-0,18	0,01	0,28	0	0
1	0,06	0,06	0,1	0,38	0,01	0	0	0
0,06	1	0,4	0,11	0,02	0,03	0,11	0	0
0,06	0,4	1	0,1	0,15	0,05	-0,08	-0,01	-0,01
0,1	0,11	0,1	1	0,03	0,12	0,04	0	0
0,38	0,02	0,15	0,03	1	0	-0,03	0	0
0,01	0,03	0,05	0,12	0	1	0,19	-0,01	0
0	0,11	-0,08	0,04	-0,03	0,19	1	0	0
0	0	-0,01	0	0	-0,01	0	1	0
0	0	-0,01	0	0	0	0	0	1

APÊNDICE 4 – Lista de espécies correlatas encontradas nas classificações

Espécie testada	Posição	Resultado da classificação - Link artigo relacionado
Acinetobacter baumannii	2	Acinetobacter nosocomialis - https://www.ncbi.nlm.nih.gov/pubmed/24965800
Acinetobacter baumannii	3	Acinetobacter pittii - https://www.ncbi.nlm.nih.gov/pubmed/24965800
Agrobacterium tumefaciens	2	Agrobacterium fabrum - http://www.uniprot.org/taxonomy/176299
Anabaena sp.	2	Aphanizomenon flos-aquae - http://www.algaebase.org/search/species/detail/?species_id=30070
Bacillus amyloliquefaciens	2	Bacillus velezensis - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5247444/ - (Bacillus amyloliquefaciens, Bacillus velezensis, Bacillus siamensis, B. subtilis)
Bacillus cereus	6	Bacillus bombysepticus - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4022800/
Bacillus cereus	2	Bacillus thuringiensis - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4022800/
Bacillus cereus	2	Bacillus weihenstephanensis - https://www.ncbi.nlm.nih.gov/pubmed/9828439
Bacillus megaterium	2	Bacillus aryabhattai - http://biosciencediscovery.com/Vol3No1/Semantairay138-145.pdf
Bacillus sp.	4	Lysinibacillus fusiformis - https://www.ncbi.nlm.nih.gov/pubmed/17473269
Bacillus subtilis	2	Bacillus atrophaeus - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC404429/
Bacillus thuringiensis	3	Bacillus bombysepticus - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4022800/
Bacillus stratosphericus	3	Bacillus altitudinis - https://www.ncbi.nlm.nih.gov/pubmed/16825614
Bacteroides sp.	0	Parabacteroides sp. - Order:Bacteroidales
Bifidobacterium adolescentis	2	Bifidobacterium stercoris - https://www.ncbi.nlm.nih.gov/pubmed/24187022
Bordetella pertussis	2	Bordetella bronchiseptica - https://www.ncbi.nlm.nih.gov/pubmed/16389302
Brucella ceti	2	Brucella pinnipedialis - https://www.ncbi.nlm.nih.gov/pubmed/17978241
Burkholderia cepacia	2	Burkholderia lata - https://www.ncbi.nlm.nih.gov/pubmed/19126732

Burkholderia pseudomallei	2	Burkholderia thailandensis - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4248825/
Burkholderia sp.	3	Paraburkholderia sp. - https://www.ncbi.nlm.nih.gov/pubmed/27054671
Camellia pitardii	2	Camellia danzaiensis - http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0073053
Chlorobaculum tepidum	2	Chlorobium tepidum - https://www.ncbi.nlm.nih.gov/pubmed/20349210
Clostridium botulinum	3	Clostridium novyi - http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0107777
Clostridium botulinum	2	Clostridium sporogenes - http://aem.asm.org/content/early/2015/06/02/AEM.01159-15
Clostridium clostridioforme	2	Clostridium bolteae - https://www.ncbi.nlm.nih.gov/pubmed/27769168
Clostridium novyi	2	Clostridium haemolyticum - https://www.ncbi.nlm.nih.gov/pubmed/11411712
Comamonas testosteroni	2	Comamonas thiooxydans - https://www.ncbi.nlm.nih.gov/pubmed/20148250
Corynebacterium glutamicum	2	Corynebacterium crenatum - C. crenatum is subspecies of C. glutamicum
Elizabethkingia meningoseptica	2	Elizabethkingia anophelis - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4868968/
Enterobacter cloacae	3	Enterobacter kobei - https://www.ncbi.nlm.nih.gov/pubmed/22827309
Enterobacter cloacae	3	Enterobacter asburiae - https://www.ncbi.nlm.nih.gov/pubmed/22827309
Enterobacter cloacae	4	Enterobacter hormaechei - https://www.ncbi.nlm.nih.gov/pubmed/22827309
Enterobacter cloacae	3	Enterobacter ludwigii - https://www.ncbi.nlm.nih.gov/pubmed/22827309
Enterobacter cloacae	3	Enterobacter xiangfangensis - http://www.microbiologyresearch.org/docserver/fulltext/ijsem/64/8/2650_ijs064709.pdf?expires=1497498125&id=id&accname=guest&checksum=8DAF08E248D05F91DC29E8E6948DDD00
Enterobacter hormaechei	2	Enterobacter xiangfangensis - https://www.ncbi.nlm.nih.gov/pubmed/24824638
Enterococcus faecalis	2	Enterococcus faecium - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3613387/

Escherichia coli	2	Shigella boydii - http://www.sciencedirect.com/science/article/pii/S167202291200112X
Geobacillus kaustophilus	2	Geobacillus thermoleovorans - http://jb.asm.org/content/194/5/1239.full
Geobacillus sp.	2	Aeribacillus pallidus - https://www.ncbi.nlm.nih.gov/pubmed/19700455
Gluconobacter oxydans	2	Gluconobacter thailandicus - https://www.ncbi.nlm.nih.gov/pubmed/15486825
Hafnia alvei	2	Hafnia paralvei - https://www.ncbi.nlm.nih.gov/pubmed/19734282
Klebsiella pneumoniae	2	Klebsiella variicola - https://www.ncbi.nlm.nih.gov/pubmed/25426853
Klebsiella pneumoniae	2	Klebsiella quasipneumoniae - http://www.bacterio.net/klebsiella.html
Lactobacillus casei	2	Lactobacillus paracasei - https://www.ncbi.nlm.nih.gov/pubmed/10499296
Lactobacillus plantarum	2	Lactobacillus paraplantarum - http://www.microbiologyresearch.org/docserver/fulltext/ijsem/46/2/ijs-46-2-595.pdf?expires=1495227825&id=id&accname=guest&checksum=7DA8DCE1C8322C0F484BDE84554AFF2A
Lelliottia amnigena	2	Enterobacter hormaechei - https://standardsingenomics.biomedcentral.com/articles/10.1186/s40793-016-0134-1
Mycobacterium bovis	3	Mycobacterium mungi - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3298296/
Mycobacterium bovis	2	Mycobacterium africanum - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC497617/
Mycobacterium chelonae	2	Mycobacterium abscessus - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4550155/
Mycobacterium intracellulare	2	Mycobacterium chimaera - http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0006263
Mycobacterium intracellulare	2	Mycobacterium indicus - http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0006263
Mycobacterium tuberculosis	2	Mycobacterium africanum - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC497617/
Mycobacterium tuberculosis	2	Mycobacterium orygis - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC497617/

Mycoplasma mycoides	2	Mycoplasma capricolum - https://www.nature.com/articles/srep19081
Paenibacillus polymyxa	2	Paenibacillus jamilae - https://www.ncbi.nlm.nih.gov/pubmed/11594596
Peptoclostridium difficile	2	Clostridium difficile - https://www.ncbi.nlm.nih.gov/pubmed/23834245
Photobacterium damselae	2	Photobacterium angustum - https://www.ncbi.nlm.nih.gov/pubmed/9828416
Photobacterium phosphoreum	2	Photobacterium kishitanii - https://www.ncbi.nlm.nih.gov/pubmed/17766874
Porphyromonas sp.	3	Parabacteroides distasonis - http://www.j-evs.com/article/S0737-0806(16)30032-6/abstract
Pseudoalteromonas haloplanktis	2	Pseudoalteromonas distincta - http://www.microbiologyresearch.org/docserver/fulltext/ijsem/50/1/0500141a.pdf?expires=1497487578&id=id&accname=guest&checksum=89A3A18480E5E8B07DCB1F3A5A5113AE
Pseudomonas amygdali	2	Pseudomonas savastanoi - https://www.ncbi.nlm.nih.gov/pubmed/22805238
Pseudomonas coronafaciens	2	Pseudomonas tremae - https://link.springer.com/article/10.1007/PL00012952
Pseudomonas fluorescens	2	Pseudomonas poae - https://microbewiki.kenyon.edu/index.php/Pseudomonas
Pseudomonas fluorescens	2	Pseudomonas azotoformans - https://microbewiki.kenyon.edu/index.php/Pseudomonas
Pseudomonas fluorescens	2	Pseudomonas synxantha - https://microbewiki.kenyon.edu/index.php/Pseudomonas
Pseudomonas fluorescens	4	Pseudomonas marginalis - https://microbewiki.kenyon.edu/index.php/Pseudomonas
Pseudomonas putida	4	Pseudomonas plecoglossicida - https://microbewiki.kenyon.edu/index.php/Pseudomonas
Pseudomonas putida	3	Pseudomonas monteilii - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1914237/
Pseudomonas syringae	4	Pseudomonas viridiflava - http://aem.asm.org/content/69/5/2936.full
Pseudomonas syringae	2	Pseudomonas amygdali - http://www.uniprot.org/proteomes/UP000037821
Pseudomonas syringae	3	Pseudomonas cannabina - https://microbewiki.kenyon.edu/index.php/Pseudomonas

Pseudomonas syringae	3	Pseudomonas avellanae - https://microbewiki.kenyon.edu/index.php/Pseudomonas
Pseudomonas syringae	6	Pseudomonas caricapapayae - https://microbewiki.kenyon.edu/index.php/Pseudomonas
Ralstonia sp.	2	Cupriavidus necator - http://www.microbiologyresearch.org/docserver/fulltext/ijsem/54/6/2285.pdf?expires=1497488335&id=id&accname=guest&checksum=2DF8E7D53807AF884FCDCA4E3CF69752
Raoultella ornithinolytica	2	Raoultella planticola - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4443617/
Rhizobium leguminosarum	2	Rhizobium acidisoli - https://www.ncbi.nlm.nih.gov/pubmed/26530784
Rhodococcus rhodochrous	2	Rhodococcus ruber - https://www.ncbi.nlm.nih.gov/pubmed/19688376
Rhodococcus rhodochrous	2	Rhodococcus aetherivorans - https://www.ncbi.nlm.nih.gov/pubmed/15053322
Rhodovulum sp.	2	Confluentimicrobium sp. - https://www.ncbi.nlm.nih.gov/pubmed/25150887
Roseburia sp.	2	Eubacterium ramulus - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC165216/
Roseovarius sp.	2	Pelagibaca bermudensis - https://www.ncbi.nlm.nih.gov/genome/?term=Pelagibaca%20bermudensis
Ruminococcus sp.	2	Blautia wexlerae - https://www.ncbi.nlm.nih.gov/pubmed/18676476s
Salmonella enterica	3	Salmonella bongori - https://www.ncbi.nlm.nih.gov/pubmed/12511502
Shigella flexneri	3	Escherichia sp. - http://www.sciencedirect.com/science/article/pii/S167202291200112X
Sphingomonas sp.	2	Novosphingobium resinovorum - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC126562/
Sphingomonas sp.	5	Sphingobium indicum - https://www.ncbi.nlm.nih.gov/pubmed/16166696
Staphylococcus aureus	2	Staphylococcus argenteus - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4898209/
Streptococcus pneumoniae	6	Streptococcus pseudopneumoniae - http://jcm.asm.org/content/44/3/923.full

Streptococcus pneumoniae	2	Streptococcus tigurinus - Streptococcus tigurinus closely related to Streptococcus mitis, Streptococcus oralis, Streptococcus pneumoniae, Streptococcus pseudopneumoniae and Streptococcus infantis - https://www.ncbi.nlm.nih.gov/pubmed/25483862
Streptococcus uberis	2	Streptococcus hongkongensis - https://www.researchgate.net/publication/233977537_Streptococcus_hongkongensis_sp_nov_isolated_from_a_patient_with_infected_puncture_wound_and_marine_flatfish
Tannerella sp.	2	Coprobacter fastidiosus - Highly possible http://www.microbiologyresearch.org/docserver/fulltext/ijsem/63/11/4181_ijs052126.pdf?expires=1497489672&id=id&accname=guest&checksum=E1E451A5D48334C11D681532D9406197
Thermoanaerobacterium saccharolyticum	2	Thermoanaerobacterium aotearoense - https://www.researchgate.net/publication/270754193_Thermoanaerobacterium_aotearoense_sp_nov_a_Slightly_Acidophilic_Anaerobic_Thermophile_Isolated_from_Various_Hot_Springs_in_New_Zealand_and_Emendation_of_the_Genus_Thermoanaerobacterium
Vibrio cholerae	2	Vibrio albensis - http://www.sciencedirect.com/science/article/pii/S0723202085800540
Wolbachia endosymbiont	2	Wolbachia pipientis - endosymbiotic bacteria
Xanthomonas axonopodis	4	Xanthomonas citri - http://onlinelibrary.wiley.com/doi/10.1046/j.1364-3703.2004.00197.x/pdf
Xanthomonas axonopodis	4	Xanthomonas perforans - http://onlinelibrary.wiley.com/doi/10.1111/epp.12018/full
Xanthomonas campestris	3	Xanthomonas euvesicatoria - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5051562/
Xanthomonas campestris	2	Xanthomonas vasicola - https://www.ndrs.org.uk/article.php?id=019025
Yersinia pestis	4	Yersinia pseudotuberculosis - https://www.ncbi.nlm.nih.gov/pubmed/22882408