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

RODRIGO DA SILVA DO NASCIMENTO

IDENTIFICAÇÃO DE PROTEÍNAS RIBOSSOMAIS EM ESPECTRO DE MASSA DO TIPO MALDI-TOF

PONTA GROSSA
2019

RODRIGO DA SILVA DO NASCIMENTO

IDENTIFICAÇÃO DE PROTEÍNAS RIBOSSOMAIS EM ESPECTRO DE MASSA DO TIPO MALDI-TOF

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.

Orientador: Prof. Dr. Arion de Campos Junior
Coorientador: Pro. Dr. Rafael Mazer Etto

PONTA GROSSA
2019

Nascimento, Rodrigo da Silva do
N244 Identificação de proteínas ribossomais em espectro de massa do tipo Maldi-Tof / Rodrigo da Silva do Nascimento. Ponta Grossa, 2019.
74 f.

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

Orientador: Prof. Dr. Arion de Campos Junior.
Coorientador: Prof. Dr. Rafael Mazer Etto.

1. Classificador glm. 2. Algoritmo genético. 3. Proteínas ribossomais. 4. Espectros de massa. 5. Modelos estatísticos. I. Junior, Arion de Campos. II. Etto, Rafael Mazer. III. Universidade Estadual de Ponta Grossa. Computação para Tecnologias em Agricultura. IV.T.

CDD: 005.3

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

TERMO DE APROVAÇÃO

Rodrigo da Silva do Nascimento

IDENTIFICAÇÃO DE PROTEÍNAS RIBOSSOMAIS EM ESPECTRO DE MASSA DO TIPO MALDI-TOF

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

Prof. Dr. Rafael Mazer Etto
Presidente

Prof. Dr. Leonardo Magalhães Cruz
UFPR

Prof. Dr. José Carlos Ferreira da Rocha
UEPG

Ponta Grossa, 11 de setembro de 2019.

Dedico este trabalho aos meus pais Julio e Maria e a minha irmã Hillary, por serem meu alento e fontes de minha dedicação, a minha namorada Ana Paula que sempre me apoiou e esteve ao meu lado e a todos que nunca deixaram de me incentivar.

AGRADECIMENTOS

Os agradecimentos principais são direcionados à Deus pelos dons e todo seu amor a mim concedidos, aos meus pais Maria de Jesus Gomes da Silva e Júlio Cardoso do Nascimento por todo amor e conselho que me deram e a minha irmã Hillary de Jesus da Silva Nascimento.

Agradeço ao meu Tio Carsimiro Gomes da Silva (in memorian) que durante sua vida me incentivou e a aconselhou-me a alcançar meus objetivos. Agradeço a Ana Paula da Silva Bertão que desde do início do mestrado me incentivou e aconselhou-me ajudando como pode mesmo com todos os trabalhos que possuía, doava de seu tempo para me ajudar, sempre estando presente e compartilhando dos meus bons e maus momentos, obrigado por estar ao meu lado.

Agradeço ao meu orientador Prof. Dr. Arion de Campos Junior, por todo seu tempo cedido, conhecimento repassado e toda orientação.

Agradeço muito ao meu coorientador Prof. Dr. Rafael Mazer Etto e a Prof a Dr a Carolina Weigert Galvão, por me conduzir nesta jornada do mestrado e por mostrar-me esse novo horizonte de conhecimentos.

Agradeço muito ao M.e Douglas Tomachewski por disponibilizar de seu tempo e seus dados para a continuação de sua pesquisa e ao Laboratório de Biologia Molecular Microbiana (LABMOM) por me acolher e compartilhar de suas dependências, para que pudéssemos realizar a pesquisa.

Agradeço muito ao Prof. Dr. José Carlos da Rochas pelas reuniões e orientações a mim concedidas, seus conhecimentos e sua prestabilidades formou o conhecimento chave para a execução desta pesquisa.

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 agradeço pelo compartilhamento de conhecimentos e dedicação à docência do programa.

Agradeço ao meu amigo Renann Rodrigues da Silva por todo incentivo e apoio.

Aos meus amigos do Lab14, aos meus colegas de Mestrado e do INFOAGRO, obrigado.

Por fim, agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo financeiro, o qual foi de extrema importância.

RESUMO

As proteínas ribossomais são utilizadas como biomarcadores na identificação taxonômica das bactérias encontradas nas subunidades 50S e 30S, a partir de sua expressão proteica no espectro de massa do tipo MALDI-TOF. É possível determinar o táxon bacteriano, com o confronto entre os espectros conhecidos e os desconhecidos. Possuindo aplicações em diversas áreas, como por exemplo, na agronomia na qual demandam da identificação de novos microrganismos que auxiliem no crescimento vegetal. Desta forma a identificação das proteínas ribossomais a partir das distribuições estatísticas se fez necessária. Para assim obter um modelo de reconhecimento probabilístico das possíveis proteínas ribossomais. Adiante foi necessário construir um Modelo Linear Generalizado - GLM, acompanhado do Algoritmo Genético, que se destaca pela simplicidade, robustez em solucionar problemas binários complexos e na seleção das variáveis para treinamento. Para a distinção dos melhores modelos classificatórios, foi levado em consideração o menor critério de informação de akaike - AIC. Os resultados foram 59 histogramas de proteínas ribossomais. Assim foram selecionadas oito distribuições estatísticas para o teste e um modelo mistura Laplaciano, algumas distribuições foram promissoras quando analisados pelos testes estatísticos e ainda assim, alguns dados se mostraram com diversos picos. O algoritmo genético mostrou-se favorável na busca de combinações de proteínas para o treinamento do modelo classificatório. Estas combinações geraram dois modelos, o primeiro com uma melhor acurácia e o outro mais específico com uma menor acurácia. A metodologia elaborada neste estudo apresentou como alternativa a discriminação de proteínas ribossomais a partir do classificador GLM.

Palavras-chave: Classificador GLM; Algoritmo Genético; Proteínas Ribossomais; Espectros de Massa; Histogramas de proteínas; Modelos Estatísticos

ABSTRACT

Ribosomal proteins are used as biomarkers in the taxonomic identification of bacteria found in the 50S and 30S subunits based on their protein expression in the MALDI-TOF mass spectrum. It is possible to determine the bacterial taxon, with the clash between known and unknown spectra. Having applications in several areas, such as agronomy in which they require the identification of new microorganisms that assist in plant growth. Thus the identification of ribosomal proteins from the statistical distributions was necessary. To obtain a probabilistic recognition model of possible ribosomal proteins. Ahead it was necessary to build a Generalized Linear Model - GLM, accompanied by the Genetic Algorithm, which stands out for its simplicity, robustness in solving complex binary problems and in the selection of variables for training. For the distinction of the best classification models, the lowest information criterion of Akaike - AIC was taken into account. The results were 59 histograms of ribosomal proteins. Thus eight statistical distributions were selected for the test and one Laplacian mixture model, some distributions were promising when analysed by the statistical tests and yet some data showed with several peaks. The genetic algorithm was favourable in the search of protein combinations for the training of the classificatory model. These combinations generated two models, the first with better accuracy and the other more specific with lower accuracy. The methodology elaborated in this study presented as an alternative to the discrimination of ribosomal proteins from the GLM classifier.

Keywords: GLM Classifier; Genetic Algorithm; Ribosomal proteins; Mass spectra; Protein histograms; Statistical Models

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE SIGLAS

AIC	Critério de Informação de Akaike
AG	Algoritmo Genético
AE	Algoritmo Evolutivo
BPCV	Bactérias Promotoras de Crescimento Vegetal
DNA	Ácido desoxirribonucleico
EM	Espectrometria de Massa
GLM	Modelo Linear Generalizado
MALDI TOF	Matrix-assisted Dessortion/Ionization time-of-flight
MS	Mass Spectrometry
M/Z	Carga/Massa
NCBI	Centro Nacional de Informação Biotecnológica
kDa	Kilodaltons
PAM	Partitioning Around Medoids
PMF	Peptide Mass Fingerprint
RNA	Ácido ribonucleico

SUMÁRIO

1 INTRODUÇÃO

A identificação bacteriana possui aplicações em diversas áreas, como por exemplo na agronomia. As identificações de espécies dos microrganismos podem contribuir em diversos aspectos, principalmente para o desenvolvimento de plantas, estes organismos são conhecidos como (BPCV) - Bactérias Promotoras de Crescimento Vegetal. Estas bactérias possuem vantagens em aplicações agrícolas como de não provocarem danos ao meio ambiente, favorecer a sanidade das plantas, protegendo-as de estresses e doenças (NOGUEIRA LONDE; SALES DIAS; SOARES DA SILVA, 2015). Assim como garantir a manutenção do ciclo biogeoquímico das culturas (HAWLEY et al., 2017). Adaptando-se a demanda moderna de alimentação, uma vez que o aumento da população vai de encontro com a busca de alimentos subsequente a expansão das terras cultivadas (CHAMIZO et al., 2018).

A diversidade de microrganismos encontrados na rizosfera das plantas podem variar, mas estima-se que 1g (um grama) de solo pode conter "2 X 10^3" a "8,3 X 10^6" de espécies bacterianas (GANS, 2005; SCHLOSS; HANDELSMAN, 2006). Estas espécies contribuem para o desenvolvimento vegetal, através do contato direto com as raízes das plantas, este processo é realizado quando estes microrganismos facilitam a degradação e à fixação dos nutrientes (BARDGETT; DE VRIES; VAN DER PUTTEN, 2017).

Duas metodologias podem ser utilizadas na identificação desses microrganismos: o método morfológico que se baseia no agrupamento de diferentes formas e arranjo das bactérias, sendo os mais comuns os cocos (forma esférica), os bacilos (forma de bastonete), os espirilos (forma espiralada) e os vibriões (forma de vírgula) (THAJUDDIN, 2018); O genotípico por meio da análise do (DNA) – Ácido Desoxirribonucleico, uma vez que os dados moleculares quando estimulados apresentam uma única dimensão, passíveis de relações matemáticas e pouco suscetíveis às condições ambientais (WOESE; KANDLER; WHEELIS, 1990). As Técnicas como a Espectrometria de Massa (EM) do tipo MALDI-TOF (matriz assistida por dessorção e ionização por tempo de voo, do inglês *Matrix-assisted Dessortion/Ionization time-of-flight*), são utilizados na identificação de organismos, isso se dá pela facilidade na preparação de amostras, sem necessidade da extração do material genético. Este método possui vantagens como processamento rápido das amostras e baixo custo o que influencia positivamente na obtenção dos resultados quando comparados aos métodos tradicionais (RODRIGUES; PASSET; BRISSE, 2018). Utilizando esta técnica são obtidos os espectros, ou seja, o perfil digital de cada organismo submetido a análise no MALDI-TOF MS, posteriormente é necessário o uso de ferramentas computacionais para interpretação dos resultados. O Ribopeaks

é uma das ferramentas mais sofisticadas para identificação de bactérias, pois utilizam as proteínas ribossomais como biomarcadores na predição (PSAROULAKI; CHOCHLAKIS, 2018).

Para identificação dos espectros, modelos de predição com suporte nos biomarcadores são utilizados na identificação taxonômica. Contudo o modelo de predição do Ribopeaks só identifica os espectros que possuem somente proteínas ribossomais com baixa taxa de erro. Este projeto tem como proposito determinar os picos ribossomais a partir do espectro de massa, pois a posição dos picos pode sobrepor uns aos outros, além de não possuírem uma posição fixa no espectro.

Em sequência são apresentados o Objetivo deste trabalho na Seção 1.1, seguidos da Revisão da literatura na seção 3, no qual será abordado a técnica de EM do tipo MALDI-TOF e seu funcionamento, bem como ocorre a atividade de identificação do resultado do MALDI e as duas metodologias computacionais para isso. Na seção 4 são apresentados a aquisição dos dados, testes estatísticos para identificação da distribuição probabilística adequada e a construção do modelo. Na seção 5 serão expostas as conclusões e possíveis trabalhos futuros.

1.1 OBJETIVOS GERAL

Criar um modelo classificatório e representativo para discriminar as proteínas Ribossomais a partir do espectro de massa MALDI-TOF.

1.2 OBJETIVOS ESPECÍFICOS

- Identificar e selecionar as proteínas ribossomais a partir do conjunto de carga massa do espectro de massa -EM;
- Construir um modelo de mistura para determinar os picos ribossomais a partir da sequência de pesos da massa carga;
- Elaborar modelos classificatórios para a predição das proteínas ribossomais a partir da base de dados do Ribopeaks.

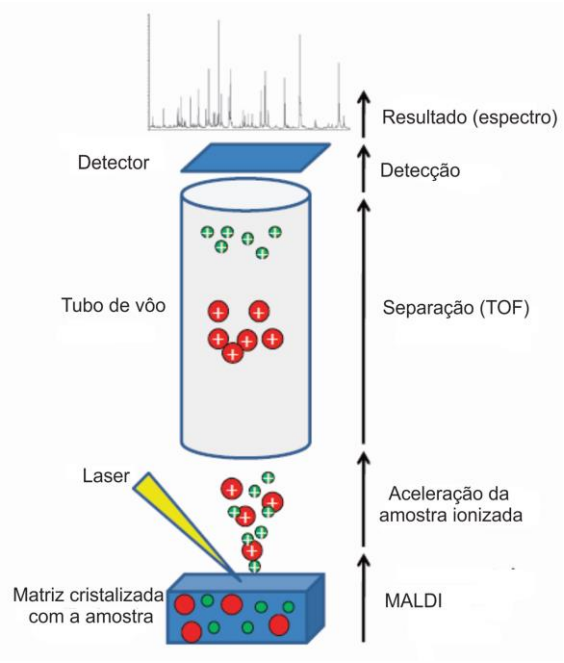
2 REVISÃO DA LITERATURA

2.1 IDENTIFICAÇÃO DE BACTÉRIAS UTILIZANDO ESPECTROMETRIA DE MASSA DO TIPO MALDI-TOF

As bactérias encontradas entre o solo e as raízes das plantas são conhecidas como rizobactérias, e algumas promovem o desenvolvimento das plantas e atuam em diferentes atividades como: metabolismo, antibiótico, resistência sistemática induzida e enzimas conferindo proteção contra patógenos e fatores de estresse ambiental (NOGUEIRA LONDE; SALES DIAS; SOARES DA SILVA, 2015; SHAMEER; PRASAD, 2018). Além de serem capazes de agir como biofertilizantes, estas bactérias necessitam ser identificadas e testadas para verificar sua interação positiva com a cultura, desempenhando um papel vital na produtividade e sustentabilidade do solo, ao mesmo tempo que protegem o ambiente (YADAV; SARKAR, 2019).

No processo de identificação das bactérias, a técnica de MS do tipo MALDI-TOF, é a mais recomendada, garante resultados rápidos e precisos de baixo custo e confiável, pois para esta técnica não há necessidade de extração do material biológico, minimizando o trabalho laboratorial (GEN; LIN, 2007; LIYANAGE et al., 2019). O resultado da análise é exibido em espectrometria, onde é possível distinguir as moléculas em relação a carga/massa (*m/z*). Esta técnica consiste no deposito da amostra biológica na matriz, que é atingida por prótons ou H^+, durante o processo de ionização das amostras ocorre o redirecionamento por um tubo de vácuo até atingir o detector ilustrado na (FIGURA 1), durante o voo as moléculas chegam em diferente tempos, as mais pesadas demoram a chegar no final do tubo enquanto as mais leves chegam primeiro, separadas de acordo com suas (*m/z*) ilustrado na (FIGURA 2) (GOULART; RESENDE, 2013). Já a sigla *m/z*, designa a quantidade de *daltons* obtido pela divisão de massa do íon pelo número de carga. A partir de células bacterianas inseridas no equipamento e processado, é gerado um espectrômetro, um perfil de expressão proteica da célula, correspondente a bactéria (PANDEY et al., 2019).

FIGURA 1- Descrição do funcionamento do MALDI-TOF

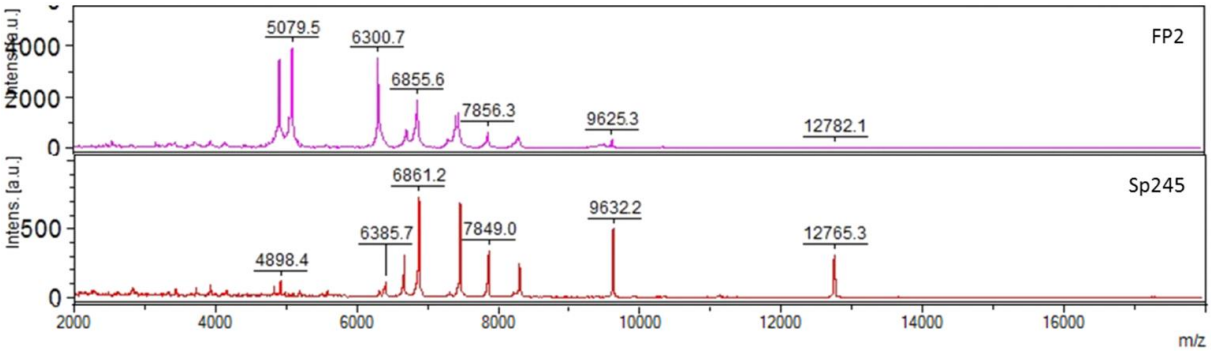


Fonte: (GOULART; RESENDE, 2013)

A partir dessas informações a identificação bacteriana ocorre de duas formas: O primeiro método necessita do espectros de Maldi conhecidos e armazenados em uma base de dados, desta forma ferramentas computacionais, são utilizados para confrontá-los e assim ocorrer a identificação do táxon (WOLK; CLARK, 2018). Para este processo existem alguns programas computacionais que utilizam essa metodologia são eles VITEK Mass Spectrometry (SAMPEDRO; CEBALLOS MENDIOLA; ALIAGA MARTÍNEZ, 2018; WOLK; CLARK, 2018), entre outros; A segunda metodologia utiliza os biomarcadores, associados a picos de massa conservados que não sofrem alteração, sem necessidade de cultivo prévio, é o caso do rRNA ribossomal 16S (NG, 2018). Alguns programas já utilizam essa metodologia como o Ribopeaks (TOMACHEWSKI et al., 2018), Mass-Up (LÓPEZ-FERNÁNDEZ et al., 2015), entre outros. Cada espectro de bactéria é conhecida como PMF – *Peptide Mass Fingerprint* única de cada microrganismos (VILLACORTA et al., 2015).

Na FIGURA 2 observamos dois espectros de massas distinto de duas bactérias com estipe diferentes, observamos que existe pequenas similaridades entre os espectros, diferenciando somente as intensidades do espectro, essas intensidades podem ser associadas a padrões de comportamento/resposta, tais informações podem ser usados para discriminar a estirpe, espécie ou gênero de uma bactéria (SILVA, 2014).

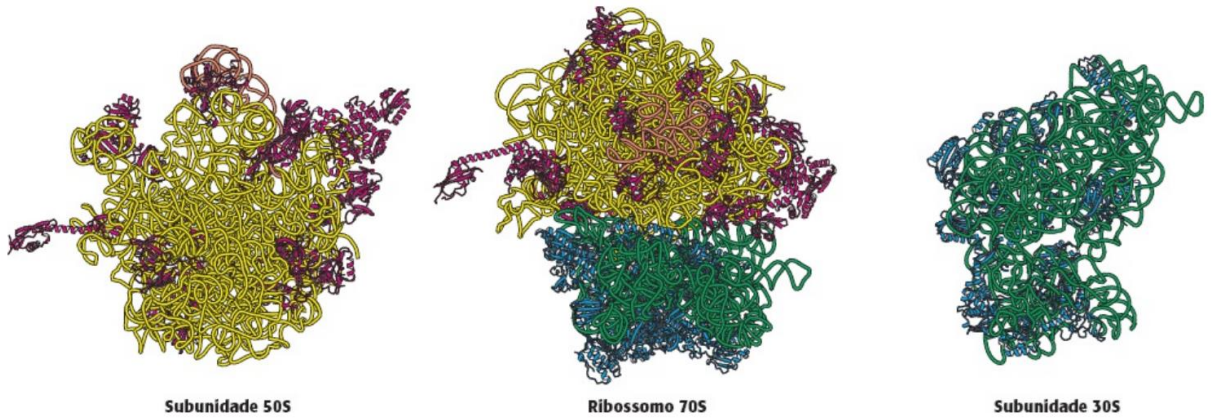
FIGURA 2- Dois diferentes tipos de espectro de massa Azospirillum brasilense FP2 e Sp245 mostrando a diferença entre as espécies



Fonte: (SILVA, 2014)

A proteína Ribossômica encontrada no Ribossomo é dividido em duas unidades: A subunidade maior (50S) contém 34 proteínas diferentes (L1 a L34) e possuem duas moléculas de RNA uma 23s e outra 5s; A subunidade menor (30s) contém 21 proteínas diferentes (designadas como S1 a S21) e uma molécula 16s. Formando desta forma o ribossomo procarioto (70s) completo, que compreende um grande complexo de nucleoproteínas ilustrado na (FIGURA 3) (BERG; TYMOCZKO; STRYER, 2014).

FIGURA 3- Modelos ribossômicos: À esquerda vista da subunidade maior (50s); À direita vista da subunidade menor (30s); No centro a formação completa do ribossomo 70s e suas interações



Fonte: (BERG; TYMOCZKO; STRYER, 2014)

2.1.1 Modelos de predição

A aprendizagem computacional é o ajuste de modelos estatísticos, moldados a partir da extração de conhecimento da base de dados, utilizado para prever uma determinada classe, baseado em variáveis independentes empregados no processo de classificação (BALDI; BRUNAK; BACH, 2001).

A classificação tem como proposito o desenvolvimento de modelos e algoritmos para identificar a categoria de um objeto a partir de um conjunto atributos ou características. Matematicamente um classificador é uma função $f: X \Rightarrow Y$ que implementa um modelo preditivo deterministico do valor da variável categórica Y a partir dos valores das variáveis $x_1 \ldots xp \in X$ (LORENA; GAMA; FACELI, 2000). Podemos utilizar as seguintes metodologias para modelagem dos classificadores: A aprendizagem supervisionada é realizada com a ajuda de especialistas da área para extração de padrões dos dados, avaliação dos resultados e testes a partir da entrada dos dados (LORENA; GAMA; FACELI, 2000; TAN; STEINBACH; KUMAR, 2009); O método de aprendizagem não-supervisionada utiliza algoritmos inteligentes para descobrir relações e padrões (LORENA; GAMA; FACELI, 2000); Por último o aprendizado semi-supervisionado utiliza parte das duas técnicas de classificação supervisionada e não supervisionada, onde algumas informações são classificadas e algoritmos inteligentes ajustam o modelo a partir das informações disponíveis para alguns dados (OLIVIER CHAPELLE BERNHARD SCHÖLKOPF, 2006).

2.2 DISTRIBUIÇÕES ESTATÍSTICAS

As distribuições estatísticas são úteis na investigação científica, constituem a ciência de coletar, analisar e interpretar dados da melhor maneira possível, importantes em situações de incerteza experimental (CHATFIELD, 2018). Podem ser categorizadas as distribuições em duas formas: Discreta representa a ocorrência de cada valor contável do tipo inteiro e não negativos (ROSS, 2014); E contínuo descreve os possíveis conjuntos de valores com intervalos do tipo infinito e incontável (LLOYD JOHNSON; KOTS; BALAKRISHNAN, 1995).

Existem diversos tipos de distribuições estatísticas entre elas: A distribuição Gaussiana também conhecida como distribuição normal (BRYC, 2012; KRISHNAMOORTHY, 2016); Log-normal (CROW.; SHIMIZU, 2018; LOPER; SUSLOW; SCHROTH, 1984); Weibull (KRISHNAMOORTHY, 2016; RINNE, 2009); Gamma (KRISHNAMOORTHY, 2016); Laplace (FORBES et al., 2011; SAMUEL KOTZ TOMAZ J. KOZUBOWSKI, 2001); Cauchy (FORBES et al., 2011); Beta (FORBES et al., 2011; KRISHNAMOORTHY, 2016); Logistic (FORBES et al., 2011; GUPTA; KUNDU, 2010); entre outras. Para cada distribuição existe parâmetros que são utilizados no comportamento da curva da distribuição ilustrado na TABELA 1.

TABELA 1- Parâmetros de probabilidades das funções de cada distribuição estatística e do modelo de mistura Laplace

	Distribuição	Notação	Parâmetros		Função de Densidade de Probabilidade
a	Normal	Normal(μ, σ^2)	μ ∈ R, média	$\alpha^2 > 0$, variância	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$
b	Log-Normal	Lognormal(μ, σ^2)	μ ∈ (-∞, +∞)	α > 0	$\frac{1}{x\sigma\sqrt{2\pi}} \exp\left(\frac{-(lnx-\mu)^2}{2\sigma^2}\right)$
c	Weibull	Weibull(λ, k)	λ ∈ (0, +∞), scale	k ∈ (0, +∞), shape	$f(x) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & x \geq 0 \\ 0 & x < 0 \end{cases}$
d	Gamma	Gamma(k, θ)	k > 0, shape	θ > 0, scale	$\frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{\frac{x}{\theta}}$
e	Laplace	Laplace(μ, β)	μ, Média ∈ R	b > 0, shape ∈ R	$\frac{1}{2b} exp\left(-\frac{\lvert x-\mu\rvert}{b}\right)$
f	Cauchy	Cauchy(x_0, γ)	x_0,Média ∈ R	γ > 0, shape ∈ R	$\frac{1}{\pi\gamma\left[1+\left(\frac{x-x_0}{\gamma}\right)\right]}$
g	Beta	Beta(α, β)	α > 0, shape ∈ R	β > 0, shape ∈ R	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{x}$ onde, $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$
h	Logistic	Logistic(μ, s)	μ, Média ∈ R	s > 0, scale ∈ R	$\frac{e^{-\frac{x-\mu}{s}}}{s(1+e^{\frac{x-\mu}{s}})}$
i	Mistura de Laplace	Mlap(k, μ, β)	$\sum_1^k$	μ, média ∈ R; b > 0, scale ∈ R	$f(x) = \sum^k n_k Laplace(x, \mu_k, \beta_k)$

Fonte adaptado: a) (BRYC, 2012; KRISHNAMOORTHY, 2016); b) (CROW.; SHIMIZU, 2018; LOPER; SUSLOW; SCHROTH, 1984); c) (KRISHNAMOORTHY, 2016; RINNE, 2009); d) (KRISHNAMOORTHY, 2016); e) (FORBES et al., 2011; SAMUEL KOTZ TOMAZ J. KOZUBOWSKI, 2001); f) (FORBES et al., 2011); g) (FORBES et al., 2011; KRISHNAMOORTHY, 2016); h) (FORBES et al., 2011; GUPTA; KUNDU, 2010); i) (MCLACHLAN; BASFORD, 1988).

2.2.1 Modelo de Mistura

O modelo de mistura possui diversas aplicações desde áreas como a medicina, biologia, engenharias, finanças, computação entre outras, pois a complexidade dos dados muitas vezes utilizada não se pode compreender utilizando simples distribuições estatísticas, a construção deste modelo depende do comportamento dos dados que são apresentados (MCLACHLAN; BASFORD, 1988).

O modelo mistura apresenta os subgrupos existentes no conjunto de dados, a partir da identificação das subpopulações é possível observar a população geral dos dados. Cada subgrupo dos dados são associados ao modelo de mistura finita a partir de um grupo de cluster

(SCRUCCA, 2016). O trabalho de Li; MacCoss; Stephens (2010) utilizou o modelo de mistura gaussiano na identificação simulada de peptídeos, a partir de uma base conhecida organizada hierarquicamente.

Para a identificação dos subgrupos algoritmos de clusterização como o K-Means são utilizados, com cada subgrupo identificado são assumidos para descrever a distribuição medida e o valor máximo das probabilidades posteriores estimadas (EVERITT, 2014).

2.3 OTIMIZAÇÃO COMBINATÓRIA

A combinação combinatória engloba problemas que necessitam de soluções que façam o melhor uso dos recursos computacionais envolvidos, não basta ter as soluções possíveis, mas também uma solução que otimize o objetivo de interesse (MIYAZAWA; DE SOUZA, 2015). Normalmente os problemas são associados a um conjunto limitado de variáveis, pode-se realizar diversas combinações em busca da melhor solução possível. Em função desta busca é possível então construir funções com objetivo de solucionar tais problemas de otimização (GOLDBARG; GOLDBARG; LUNA, 2017).

Os problemas combinatórios podem ser de minimização ou de maximização. Nos dois casos, temos a função aplicada a um conjunto de dados finito, que em geral é enumerável. Apesar de finito, em geral a combinação de cada elemento é computacionalmente oneroso e impraticáveis com algoritmos simples que não possuem uma estratégia (DEB, 2001; GOLDBARG; GOLDBARG; LUNA, 2017; SANTOS; MARCELOS, 2016). Desta forma surgem os métodos de otimização que buscam soluções exatas ou alternativas, conhecidas como heurística próximos a solução real (GREEN; ALETI; GARCIA, 2017).

Poucos problemas combinatórios são resolvidos de forma exata, pois a busca pela solução exata possui uma complexidade muito alta, por isso muitos outros problemas são esperados algoritmos exatos e rápidos para problemas de grande porte (DONG et al., 2018).

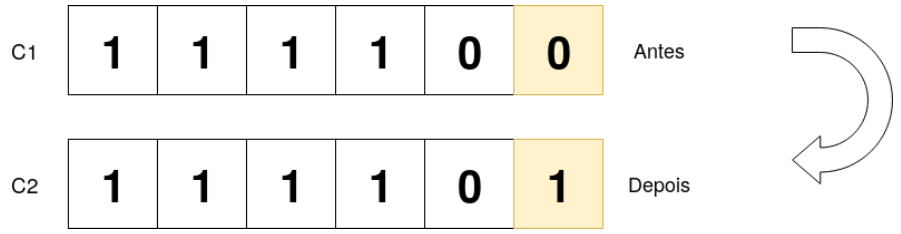
2.4 ALGORITMO GENÉTICO - AG

O Algoritmo Genético pertence ao ramo dos algoritmos evolutivos - AE na qual combina a computação e diversas áreas de conhecimento como matemática, biologia, química, física e engenharia (KUMAR; JAIN; SHARMA, 2018). Desta forma o AG baseado na teoria da evolução das espécies, onde os indivíduos por meio da seleção natural e interação entre as espécies foram capazes de adaptação ao meio ambiente, progrediram em estruturas complexas que permitem a sobrevivência em diferentes tipos de ambientes e o cruzamento transporta a

ascendência, as principais características do sucesso da evolução, o que torna o algoritmo atraente e flexível a diversos problemas de otimização (KRAMER, 2017).

As possíveis soluções do AG binário, são representadas por estruturas denominadas cromossomos ou cadeia binárias de tamanho fixo, a (FIGURA 4) exemplifica uma solução representada por um cromossomo binário, essas cadeias são tratadas como soluções e passam por mutação, seleção e recombinação por várias iterações (COELHO, 2013).

FIGURA 4- Exemplo de um cromossomo clássico de um AG binário



Fonte: (COELHO, 2013)

O pseudocódigo do algoritmo genético básico é ilustrado pelo (QUADRO 1), usado como modelo para muitas abordagens relacionadas. No início do algoritmo ocorre a inicialização aleatória da população binária para cobrir aleatoriamente todo o espaço de busca. Então é calculado a adequação dos indivíduos e selecionam somente os melhores para a próxima iteração, com base em sua adequação. Esse processo é repetido até que os critérios de término sejam atendidos (COELHO, 2013; KRAMER, 2017; UMBARKAR; SHETH, 2015).

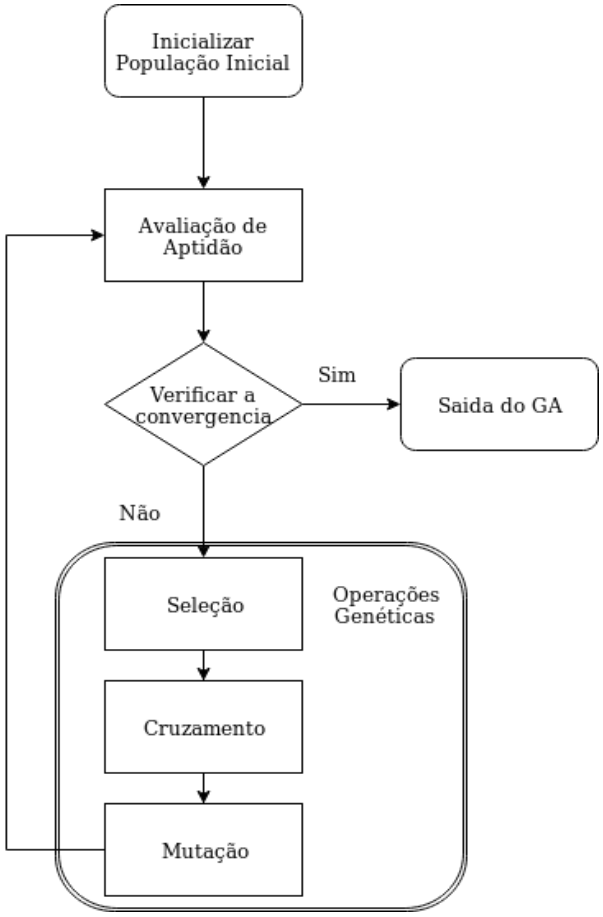
QUADRO 1- Algoritmo genético padrão

Inicializar a População
Repetir
    Repetir
        Recombinar - crossover
        Mutação - mutation
        Mapeamento fenótipico - phenotype mapping
        Aptidão computacional - fitness computation
    até população completa
    seleção de população parental
até condição de termino

Fonte: (KUMAR; JAIN; SHARMA, 2018)

O GA é um dos algoritmos evolutivos de maior sucesso baseado na seleção natural, usado para se isentar de problemas de otimização multifacetados utilizando operações com inspiração biológica, passando a impressão do processo de evolução, a FIGURA 5 mostra os principais passos seguidos pelo GA (KUMAR; JAIN; SHARMA, 2018).

FIGURA 5- Fluxograma do algoritmo genético

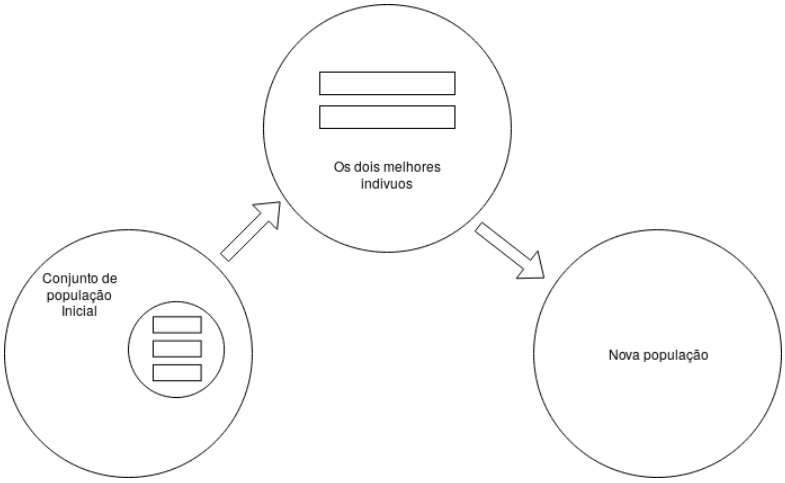


Fonte: (KUMAR; JAIN; SHARMA, 2018)

2.4.1 Seleção

A seleção é o primeiro passo na escolha de indivíduos para reprodução. A escolha dos indivíduos é feita de forma aleatória a partir de probabilidades, com o intuito de os melhores indivíduos sejam selecionados para reprodução (SIVANANDAM; DEEPA, 2007). De acordo com a função de aptidão, maior a chance de um indivíduo ser selecionado, desta forma a seleção leva o GA a melhorar a aptidão da população nas gerações sucessivas (KUMAR; JAIN; SHARMA, 2018; SIVANANDAM; DEEPA, 2007). A FIGURA 6 ilustra o funcionamento básico do processo de seleção.

FIGURA 6- Seleção dos indivíduos

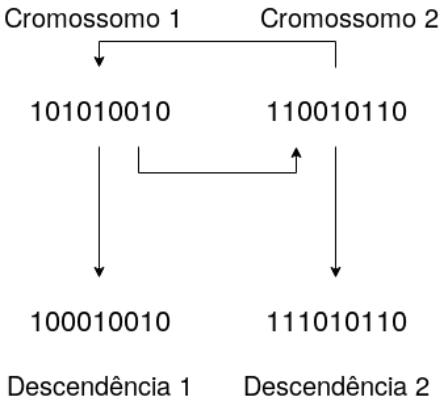


Fonte: (SIVANANDAM; DEEPA, 2008)

2.4.2 Cruzamento

O cruzamento é uma operação que permite a combinação do material genético de duas ou mais soluções (SYSWERDA, 1993). Na qual duas cadeias de bits com tamanho fixo usado como pais, geram novos indivíduos como ilustrado na FIGURA 7 (KRAMER, 2017; SIVANANDAM; DEEPA, 2007). Ao final da execução, o cromossomo com melhor cruzamento é definido como solução ótima do problema.

FIGURA 7- Exemplo de cruzamento entre dois cromossomos



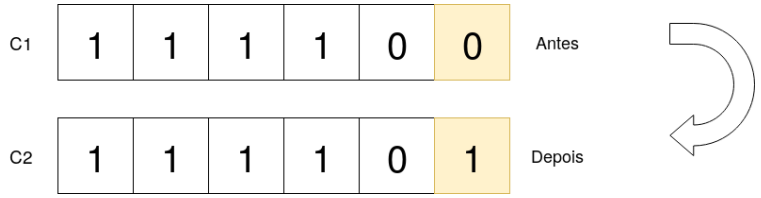
Fonte: (UMBARKAR; SHETH, 2015)

2.4.3 Mutação

A mutação é uma mudança aleatória, afim de evitar que o AG gere indivíduos parecidos e uma consequente perda nos novos indivíduos, permitindo desta forma que o AG busque por novas soluções, além de impedir que o algoritmo fique preso no mínimo local (SIVANANDAM; DEEPA, 2007). O operador de mutação troca o gene do cromossomo, a partir

de uma determinada probabilidade, geralmente menor que 1%, como ilustrado na FIGURA 8 o operador troca a posição do gene do cromossomo (AURÉLIO CAVALCANTI PACHECO, 1999). Se o cruzamento ajuda na exploração da solução atual e assim determinar os melhores indivíduos, a mutação ajuda na exploração de todo o espaço de busca.

FIGURA 8- Exemplo de mutação do gene



Fonte: (COELHO, 2013)

A partir da mutação a diversidade genética se mantem entre a população, pois introduz novas estruturas genéticas na população modificando somente alguns de seus blocos de construção. Além de manter o pool genético para garantir a ergodicidade (COELHO, 2013; SIVANANDAM; DEEPA, 2007).

2.5 MODELO LINEAR GENERALIZADO – MLG

O modelo linear Generalizado (do inglês *Generalized Linear Model – GLM*) é uma generalização do modelo linear, que proporciona uma variável resposta a partir de modelos que não seguem uma distribuição normal (DOBSON; BARNETT, 2008). Assume-se que a variável reposta Y de interesse primário, e um vetor $\mathbf{X} = (x_1, \ldots, x_k)^T$ de k variáveis explicativas, que esclarece parte da variabilidade inerente a Y. A generalização ocorre devido a função de ligação que transforma a variável dependente, sendo mais usuais a logística, probit e complemento log-log (SOUZA, 2015). A variável resposta a Y pode ser contínua, discreta ou dicotómica. Esse modelo envolve uma variável resposta univariada, variáveis explanatórias e uma amostra aleatória de n observações independentes (RODRIGUEZ, 2016).

Assim, um MLG é definido por uma distribuição de probabilidade, para a variável resposta, um conjunto de variáveis independentes descrevendo a estrutura linear do modelo e uma função de ligação entre a média da variável resposta e a estrutura linear (CORDEIRO; DEMÉTRIO, 2008).

2.6 TRABALHOS CORRELATOS

Shen et al. (2008) desenvolveram um modelo estatístico hierárquico (HSM) que modela conjuntamente a incerteza dos peptídeos e proteínas

Utilizando modelos estatísticos e dados do MALDI-TOF do tipo MS, foram aplicados técnicas computacionais de aprendizagem, conhecidas como máquinas de vetores de suporte e florestas aleatórias, resultaram na precisão de 94% a 98% na identificação de bactérias no trabalho de (DE BRUYNE et al., 2011). Os autores Santos (2017), Tomachewski et al. 2018, Ziegler et al. (2015) utilizaram biomarcadores ribossomais como picos chaves na identificação taxonômica das bactérias promotoras de crescimento vegetal na criação dos modelos.

Ziegler et al. (2015) relata que o modelo de aprendizagem computacional foi eficaz em 116 estirpes nos gêneros rizóbios e 13 proteínas ribossômicas foram suficientes na identificação confiável das bactérias. Com este estudo montou-se uma base de dados com todas as informações ribossômicas extraídas de amostras biológicas, específicos para promover o crescimento vegetal conhecidos por nodular raízes de plantas, gerando nódulos que fixam o nitrogênio da atmosfera.

Santos (2017) utilizou método de imputação, na qual contorna a falta de informação na base de dados ribossomal, utilizando o método (K-NN) *K-nearest neighbors* com a média, mediana e zero para imputação dos dados, somente em casos que houvesse de um há três atributos faltantes, posteriormente o modelo ajustado era aplicado a regressão logística binária. A metodologia foi aplicada na base de dados conhecida como PUKYU Galvão et al. (2017), ajustada com os grupos bacterianos mais populosos, foram eles *bacterium abscessus* e *Mycobacterium fortuitum*.

O trabalho de Tomachewski (2017) desenvolveu a base de dados PUKYU por meio da aquisição dos dados obtidos do repositório *National Center Biotechnology International* (NCBI), selecionaram as proteínas ribossomais e calcularam os pesos moleculares procurando por marcadores que poderiam identificar um microrganismo, utilizaram as proteínas ribossomais como biomarcadores por possuírem alta conservação em diferentes condições fisiológicas. Em seguida utilizaram a aprendizagem de máquina não-supervisionada para criação de modelo estatístico para reconhecimento taxonômico das bactérias.

O modelo mistura foi empregado para a classificação de vegetação na região amazônica, com o objetivo de identificar as espécies em diferentes estágios de classificação, florestas secessionais, maduras intermediárias e avançadas alcançando uma acurácia geral de 78,2% (LU; MORAN; BATISTELLA, 2003).

(LI; MACCOSS; STEPHENS, 2010) desenvolveram uma abordagem estatística que a partir de um espectro de massa, verificaram quais proteínas e peptídeos estão presentes na amostra e quais proteínas estão presentes e quais peptídeos são identificados corretamente, permitindo um retorno de evidência apropriado entre as proteínas e seus peptídeos constituintes.

(MANOGARAN et al., 2018) utilizou o modelo de mistura para criação de um *framework* que auxilia-se no processo de diagnóstico de câncer, por meio do modelo Bayesiano oculto de *Markov* (HMM), mistura de distribuições gaussianas (MG), foram analisados o agrupamento dos dados da sequência de DNA em todo o genoma, possuindo o objetivo de detectar o distúrbio genético e modelar o diagnóstico da doença utilizando a ferramenta desenvolvida na pesquisa.

3 MATERIAIS E MÉTODOS

3.1 DESCRIÇAO DA BASE DE DADOS USADAS NOS EXPERIMENTOS

A base de dados utilizada no desenvolvimento desta pesquisa do tipo insilico oriundos do trabalho de Tomachewski (2017), conhecida como PUKYU, possuem atributos representativos de picos ribossomais de espectros de massas do tipo MALDI-TOF pré-identificadas. Especificamente a base de dados contém informações das seguintes proteínas: 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, S21, S22 e S31e. Para montagem da base de dados PUKYU foram obtidos a partir do repositório do NCBI – Centro Nacional de Informação Biotecnologia, todos referentes a proteínas ribossomais (30S e 50S) e cada sequência de aminoácidos foram processados na Calculadora de Peptídeos desenvolvida no projeto de Tomachewski (2017), na data de 13 de junho de 2016, por meio da API - Interface de aplicação a programação, do inglês (*Application Programming Interface*). Complementarmente a base PUKYU assume as leituras do espectro ao intervalo entre 0 a 40kDa.

Com intuito de identificar as distribuições estatísticas, contruir o modelo de mistura laplaciano e o Modelo Linear Generalizado – GLM *Generalized Linear Model* para identificação de proteínas ribossomais. Foram criados 61 novas base de dados a partir da base de dados PUKYU, cada base de dados criada e ilustrada na TABELA 2 possuem a informação dos kDa de cada proteína e a quantidade de registro para cada.

TABELA 2- Partição de proteínas ribossomais

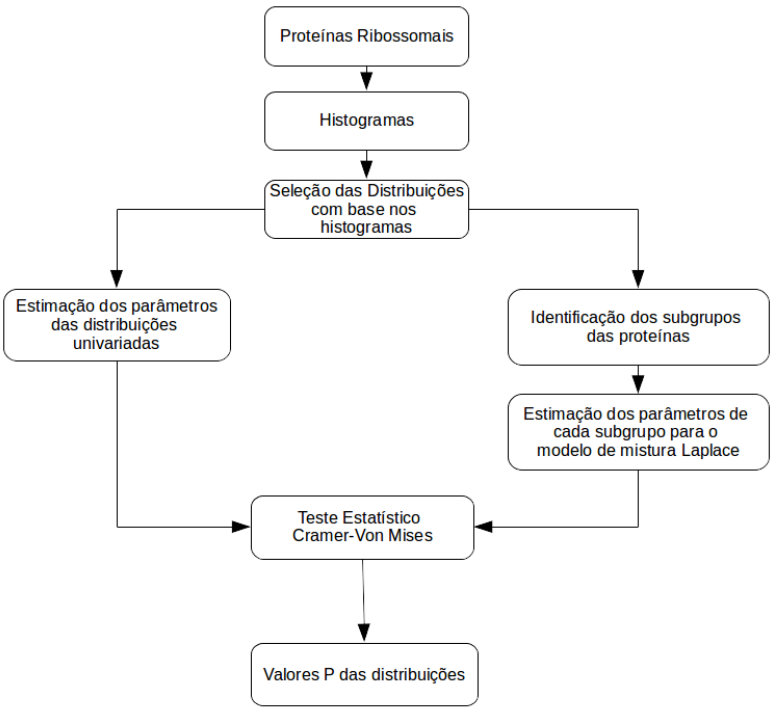
L1	L2	L3	L4	L5	L6	L7
22168	21732	26585	26409	23760	22397	2409
L7ae	L7L12	L9	L10	L11	L12	L13
4670	20365	21642	23326	21714	621	21763
L15	L16	L17	L18	L19	L20	L21
22068	21143	21470	21669	21860	21032	21696
L23	L24	L25	L27	L28	L29	L30
23531	21712	19565	21109	20548	20811	19572
L32	L33	L34	L35	L36	S1	S2
20066	20020	16542	20210	16550	1022	22164
S4	S5	S6	S7	S8	S9	S10
21898	21407	21825	2336	23741	23658	20022
S12	S13	S14	S15	S16	S17	S18
20616	25790	21450	21511	21355	25831	20452
S21	S22					
16115	541					

A principal motivação para a identificação das distribuições estatísticas é o fato de estar relacionamento ao comportamento dos dados, pois desta forma será possível correlacionar as proteínas com suas distribuições estatísticas (GALVÃO et al., 2017; LI; MACCOSS; STEPHENS, 2010; TOMACHEWSKI et al., 2018).

3.2 IDENTIFICAÇÃO DAS DISTRIBUIÇÕES DE PROTEÍNAS NA BASES DE DADOS

O pré-processamento foi o primeiro passo para iniciar a identificação das distribuições dos dados, desta forma foi utilizado a avaliação por histogramas e a verificação dos modelos. Para isso foi levado em consideração os fluxos ilustrados na Figura 9.

FIGURA 9- Fluxo da seleção das distribuições dos dados para as proteínas ribossomais



Fonte: (COX, 2017)

Em seguida foi realizado o teste estatístico, Cramer Von Mises (CVM) também conhecido como teste de significância ou teste de hipótese, permitindo tomar uma decisão entre duas hipótese nula H_0 e a alternativa H_1ilustrado abaixo .

$$\begin{cases} H_0: & Os\ dados\ seguem\ a\ distribuição \\ H_1: & Os\ dados\ n\ o\ seguem\ a\ distribuição \end{cases}$$

3.2.1 Teste Cramer-Von Miser

O teste Cramer-Von Miser é um critério que foi utilizado para julgar a qualidade do ajuste de uma função bseado na distribuição cumulativa. Utilizando no projeto a partir do pacote *goftest* pertencente a linguagem R, que utiliza a equação abaixo:

$$W = \frac{1}{12n} + \sum_{i=1}^{n} (p(i) - \frac{2i-1}{2n})$$

Esse teste foi utilizado para verificar se os dados possuiam a mesma estrutura do comportamento da distribuição estatística (ALMEIDA; DE MARTINIS, 2019; WANG et al., 2019).

3.3 DISTRIBUIÇÃO ESTATÍSTICAS

A seleção das distribuições levou em consideração o tipo de dados encontrados nas bases de dados biológicas. Esta base por ser considerada do tipo contínuo foi selecionada nove

distirbuições estatísticas, seguindo os critérios da Figura 9 são elas: Distribuição Normal; Log-normal; Weibull; Gamma; Laplace; Cauchy; Beta; Logistica e Exponential.

3.3.1 Distribuição Normal

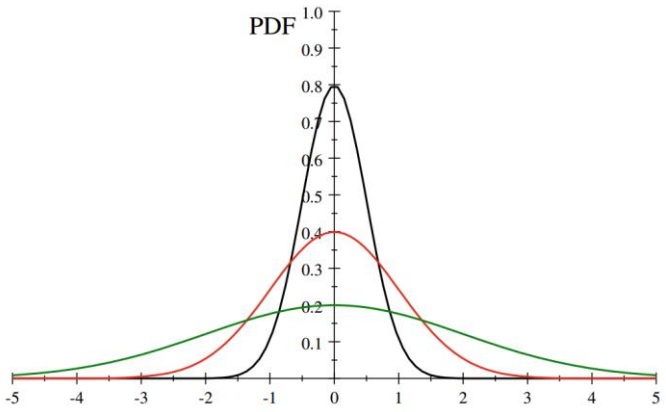
A distribuição normal é do tipo discreta e contínua também conhecida como gaussiana, a aplicação se deve a usos na modelagem de diversos fenômenos naturais (BRYC, 2012; KRISHNAMOORTHY, 2016).

A partir de um vetor X que segue a distribuição normal *N(μ, σ)* com parâmetros de média μ e desvio padrão σ se sua função densidade de probabilidade (FDP), seguir a função $f_n(x, \mu, \sigma)$ na forma da equação abaixo (AHSANULLAH, 2017):

$$f_n(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, -\infty < \mu < x < \infty, \sigma > 0$$

Na Figura 10 ilustramos o comportamento da função de densidade da distribuição normal.

FIGURA 10 -Densidade de probabilidade da distribuição normal



Fonte: (BRYC, 2012)

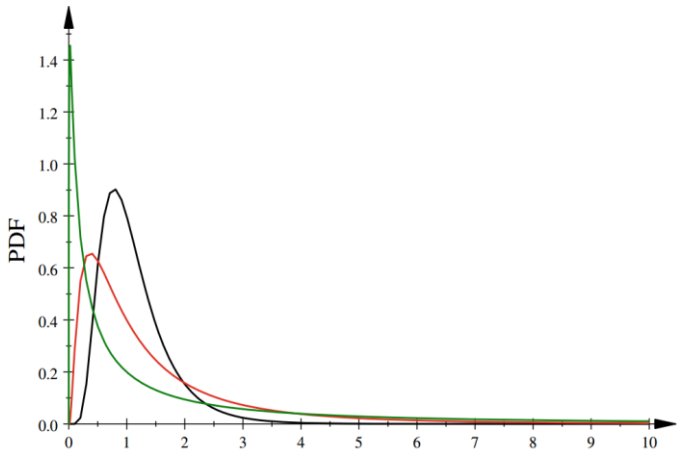
3.3.2 Distribuição Log-normal

De acordo com Ahsanullah (2017) e Famoye (1995), um vetor de dados X é denominado distribuição log-normal quando o logaritmo de seus dados $Y = \log(X)$ tem a distribuição normal e sua função de densidade segue a equação abaixo:

$$P_x(x) = \delta[(x-\theta)\sqrt{2\pi}]^{-1}\exp[\frac{-1}{2}\{\gamma + \delta\log(x-\theta)\}^2], x > \theta$$

Ahsanullah (2017) relata que os cálculos dos parâmetros são os seguintes: Média μ é $e^{\mu+\frac{\sigma^2}{2}}$; variância σ é $(e^{\sigma^2} - 1)^{e^2\mu+\sigma^2}$. O comportamento de seus dados pode ser visualizado na Figura 11.

FIGURA 11- Densidade de probabilidade da distribuição Log-normal



Fonte: (AHSANULLAH, 2017)

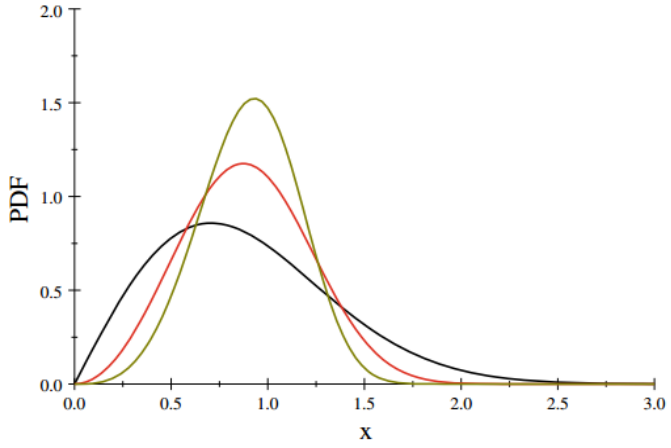
3.3.3 Distribuição Weibull

A distribuição *Weibull* é do tipo contínua, já utilizada de forma bem sucedida em dados provenientes das áreas biológicas, ambientais, engenharias, entre outras (RINNE, 2008). O vetor de dados segue a distribuição Weibull se a função densidade de seus dados $f_{wb}(x, \mu, \sigma, \delta)$ segue a equação abaixo:

$$f_{wb}(x, \mu, \sigma, \delta) = \frac{\delta}{\sigma}\left(\frac{x-\mu}{\sigma}\right)^{\delta-1} e^{\frac{x-\mu}{\sigma}}, -\infty < \mu < x < \infty, \sigma > 0, \delta > 0$$

Os cálculos dos parâmetros são os seguintes: Média μ é $\mu + \Gamma(1 + \frac{1}{\delta})$; variância$\sigma^2((\Gamma 1 + \frac{2}{\delta}) - (\Gamma(1 + \frac{1}{\sigma}))^2)$. A Figura 12 ilustra o comportamento dos dados a partir da distribuição Weibull.

FIGURA 12- Densidade de probabilidade da distribuição Weibull



Fonte: (RINNE, 2008)

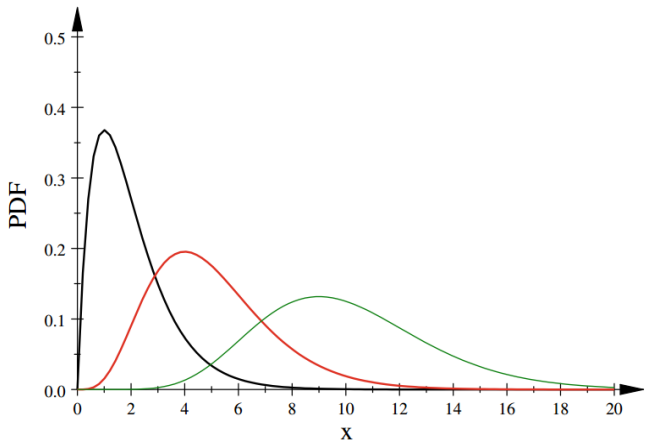
3.3.4 Distribuição Gamma

A distribuição Gamma é utilizada na modelagem de dados positivos e são assimétricos a direita e maiores que 0 e pertencem a distribuição do tipo continua, a partir de um vetor X dado os parâmetros $f_{GA}(a,b)$. Segue a distribuição *gamma* a partir da equação abaixo:

$$f_{GA}(a,b,x) = \frac{1}{\Gamma(a)b^a} x^{a-1}e^{-x/b}, x \geq 0, a > 0, b > 0$$

Os cálculos dos parâmetros são os seguintes: Média é *ab*; Variância é ab^2, e podem ser visualizados o comportamento dos dados na Figura 13.

FIGURA 13- Densidade de probabilidade da distribuição Gamma



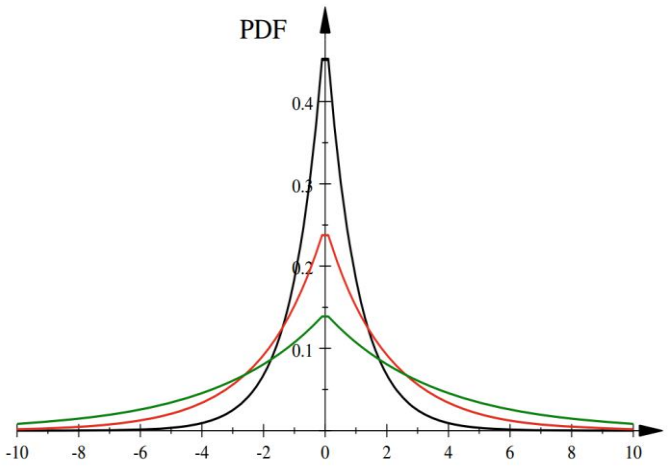
3.3.5 Distribuição Laplace

A distribuição Laplace também conhecida como exponencial dupla é utilizada quando os dados apresentar mais picos que a distribuição normal, muito utilizada na modelagem de dados biológicos, financeiros entre outros. Lloyd Johnson, Kots, Balakrishnan (1995) relata que a função de densidade da distribuição segue a equação abaixo.

$$f(x,\mu,b) = \frac{1}{2b}\exp(\frac{-|x-\mu|}{b})$$

O comportamento da distribuição é ilustrado na Figura 14, este comportamento é possível quando os dados demonstram mais de um pico.

FIGURA 14 - Densidade de probabilidade da distribuição Laplace



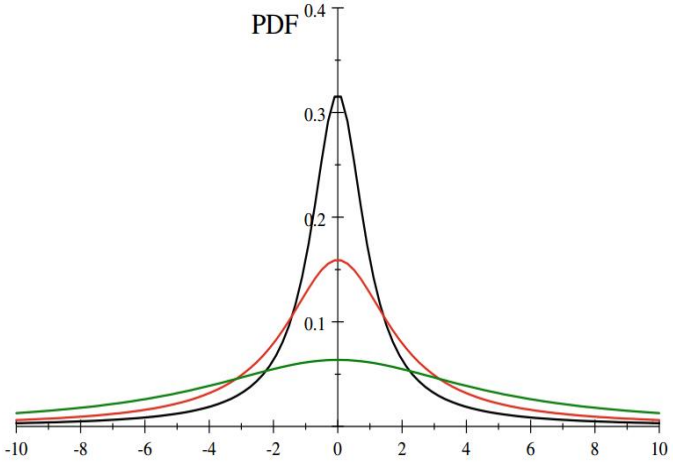
3.3.6 Distribuição Cauchy

A distribuição Cauchy é considerada uma distribuição patológica, pois não apresenta média e variância, possui importância em diversas áreas do conhecimento científico na física, matemática onde é uma das soluções para a equação de Laplace, entre outras finalidades. A função de densidade é dada pela equação abaixo.

$$f_c = \frac{1}{\pi\sigma(1+(\frac{x-\mu}{\sigma})^2)}, -\infty < x < \mu < \infty, \sigma > 0$$

O comportamento dos dados da distribuição Cauchy segue na Figura 15. A partir desta informação e possível selecionar as distribuições que possam seguir o comportamento dos dados da base PUKYU.

FIGURA 15- Densidade de probabilidade da distribuição Cauchy



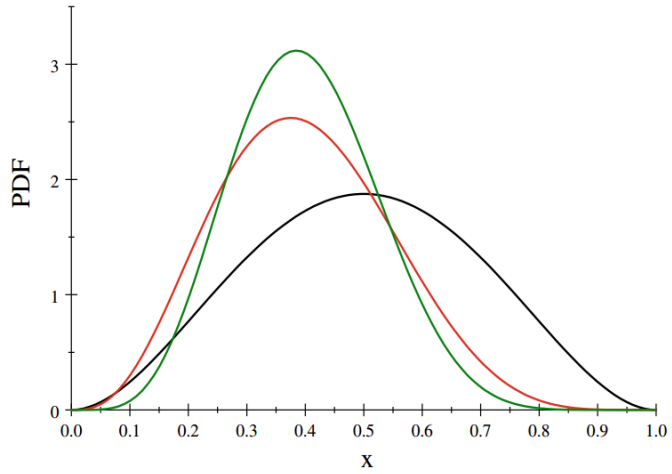
3.3.7 Distribuição Beta

A distribuição Beta do tipo continua é definida pelo intervalo [0,1] e os parâmetros da distribuição são positivos e denotados por α e β. É dito se o vetor X possui comportamento da distribuição Beta se a sua função densidade de probabilidade seguir a equação abaixo.

$$f_{mn}(x) = \frac{1}{B(m,n)} x^{m-1}(1-x)^{n-1}, 0 < x < 1, m > 0, n > 0$$

Os cálculos dos parâmetros são os seguintes: Média é $\frac{m}{m+n}$; Variância é $\frac{mn}{(m+n)^2(m+n+1)}$. O comportamento da distribução Beta é possível ser vista na FIGURA 16.

FIGURA 16- Densidade de probabilidade da distribuição Beta

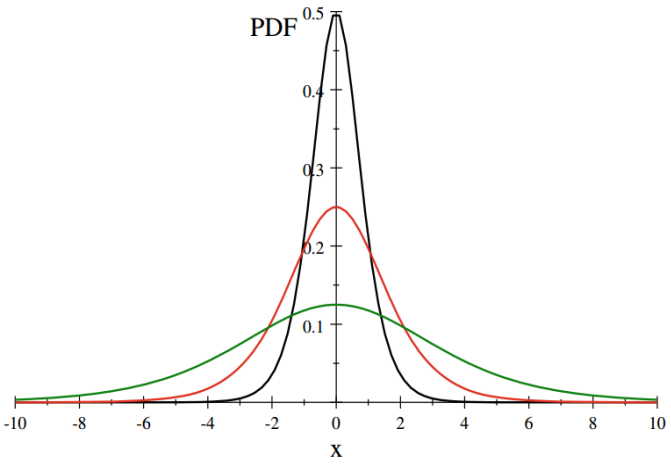


3.3.8 Distribuição Logística

A distribuição Logística é muito aplicado em áreas de processo produtivo, é digo que um vetor X pertence a distribuição logística $LG(\mu, \sigma)$com os parâmetros de média μ e parâmetro de desvio padrão σ se sua função de densidade de probabilidade seguir a equação abaixo. O comportamento dos dados pode ser visualizado na Figura 17.

$$f_{x,\mu,\sigma} = \frac{1}{\sigma} \frac{e^{\frac{-x-u}{\sigma}}}{(1+\frac{e^{-x-\mu}}{\sigma})^2}, -\infty < \mu < x < \infty, \sigma > 0$$

FIGURA 17- Densidade de probabilidade da distribuição Logistic



3.4 MODELO DE MISTURA LAPLACIANO

O modelo de mistura foi aplicado nas proteínas relatadas na seção 4.1, e a estimativa dos parâmetros pelo algoritmo de clusterização na identificação dos subgrupos. Para isso o algoritmo k-medoids foi utilizado para particionar um conjunto de dados em k cluster. Para isso foi utilizaddo o algoritmo PAM ou também conhecido K-medoids com estimador de cluster a partir dos dados.

O algoritmo PAM particiona o conjunto de dados de n objetos em k clusters, onde o conjunto de dados e o número k, são a entrada do algoritmo, trabalhando com a dissimilaridade, com o objetivo de minimiza-la entre os representantes de cada subgrupo e seus membros (LEONARD KAUFMAN, 2005). O algoritmo utiliza a seguinte função na seleção do problema $F(x) = minimize \sum_{i=1}^{n} \sum_{j=1}^{n} \quad d\,(i,j) z_{ij}$.

Onde a Função X é utilizada para minimizar, $d(i,j)$ é a medida de dissimilaridade entre as entidades i e j, e Z_{ij} é a variável que garante apenas a dissimilaridade entre entidades do mesmo cluster.

O próximo passo a partir da identificação dos subgrupos foi construir o modelo de mistura definido pela equação $f(x) = \sum_{i=1}^{n} d\, w_i g_i(x)$, Onde o peso $0 \leq w_i \leq 1$ para cada sub-grupo i = 1, 2, 3, 4, …, n da mistura w_i a soma é igual a 1.

3.5 MODELAGEM DO PROBLEMA

Para executar a seleção do melhor modelo classificatório para as proteínas ribossomais com GA, foi necessário elaborar o problema de otimização, para isso os genes do cromossomo foram elaborados como um vetor que corresponde-se a um atributo do tipo

proteínas. Optou-se pelo uso do GA pois a versão binária corresponde melhor ao problema, visto que o mesmo foi criado para problemas binários (KOZA, 1997).

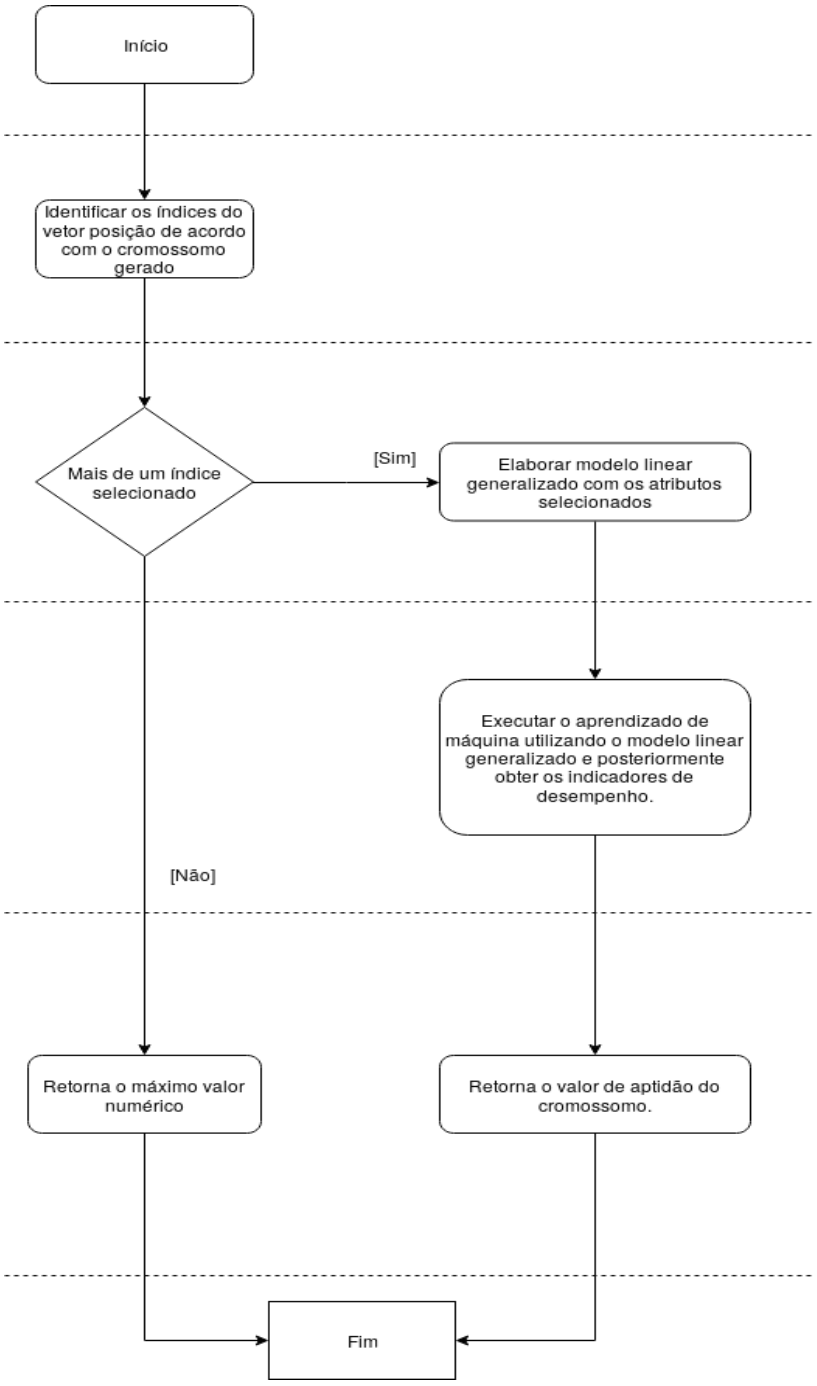
3.5.1 Codificação dos cromossomos

Os cromossomos foram codificados de forma que cada índice do vetor posição correspondeu a um atributo de entrada do conjunto de dados, portanto a dimensão dos vetores, *d*, foi igual a 59. Em virtude do limiar interferir na quantidade de atributos selecionados e que o valor possa variar conforme o conjunto selecionado para a classificação do nosso cenário foi utilizado a seguinte função $maximizar \sum_{i=1}^{n} \quad x_i$; variáveis binárias X_i (i = 1, ...n) que assumem o valor 1, caso a proteína que consta no vetor *X* seja selecionado para a construção do modelo, onde $x_i = 0 ou 1$.

3.5.2 Função de aptidão

A FIGURA 18, ilustra que no início do fluxograma é recebido o cromossomo gerado a partir do GA, na qual posteriormente são identificados os índices do vetor de posição. Em seguida é analisado a seleção do cromossomo, se for acima de 1 (um) índice selecionado, então é realizado a montagem da equação do modelo para que seja efetuado o treinamento, em seguida o valor de predição do modelo gerado. Caso o cromossomo concebido selecione somente um parâmetro do vetor é retornado o valor máximo do modelo.

FIGURA 18- Etapas executadas para a construção do modelo de classificação com Algoritmo Genético



Para avaliação dos modelos gerados foi utilizado a equação $AIC = -2(\log - likehood) + 2k$, na qual k é o número de parâmetros utilizados para treinamento, *log-likehood* é uma medida de ajuste do modelo. Quanto maior o número, melhor o ajuste, geralmente obtido na saída estatística.

3.5.3 Parâmetros do Algoritmo Genético – AG

A versão do algoritmo genético utilizada neste trabalho foi a versão binária que reuni a versão original do algoritmo com todas as melhorias (SCRUCCA, 2013). Foram investigados os valores ideias para números de gerações, tamanho da população e o valor de limiar que proporcionaram o melhor resultado do conjunto de dados. Para isso os seguintes parâmetros foram utilizados:

- **Números de iterações:** 20, 50
- **Tamanho da população:** 20, 40
- **Dimensão do cromossomo:** 59
- **Espaço de busca Inferior (Mínimo):** 0
- **Espaço de busca Superior (Máximo):** 1
- **Critério de parada:** Atingir o número máximo de iterações
- **Taxa de cruzamento:** 0,5
- **Taxa de mutação:** 0,03
- **Seleção:** Torneio
- **Mutação:** Aleatório uniforme

4 RESULTADOS E DISCUSSÃO

Foram criados histogramas de 59 proteínas ribossomais disponíveis na base de dados *PUKYU* encontrados no APÊNDICE A – Histograma das proteínas Ribossomais. Os histogramas foram desenvolvidos para selecionar as distribuições de acordo com cada histograma, seguindo as etapas ilustrado na FIGURA 9 (COX, 2017).

A partir dos histogramas foram selecionadas algumas distribuições, são elas a distribuição Normal, Log-Normal, Weibull, Gamma, Laplace, Cauchy, Beta, Logistic e o modelo mistura laplaciano. Para representar os dados que mostrassem mais de um pico, as proteínas foram clusterizadas para a identificação dos subgrupos, os resultados podem ser visualizados na TABELA 3, na qual a partir desta informação foi possível a modelagem do modelo mistura laplaciano.

TABELA 3- Agrupamento identificados para cada proteína

Proteínas	Subgrupos	Proteínas	Subgrupos	Proteínas	Subgrupos
L1	5	L20	2	S4	9
L2	2	L21	2	S5	2
L3	2	L22	2	S6	3
L4	2	L23	10	S7	2
L5	5	L24	2	S8	9
L7a	5	L25	2	S9	2
L7ae	5	L27	3	S10	2
L7L12	9	L28	4	S11	2
L9	2	L29	5	S12	3
L10	10	L30	2	S13	10
L11	2	L31	2	S14	2
L12	2	L32	2	S15	2
L13	2	L33	2	S16	2
L14	10	L34	2	S17	2
L15	2	L35	3	S18	2
L16	6	L36	4	S19	3
L17	2	S1	2	S20	8
L18	7	S2	10	S21	6
L19	2	S3	6	S22	10

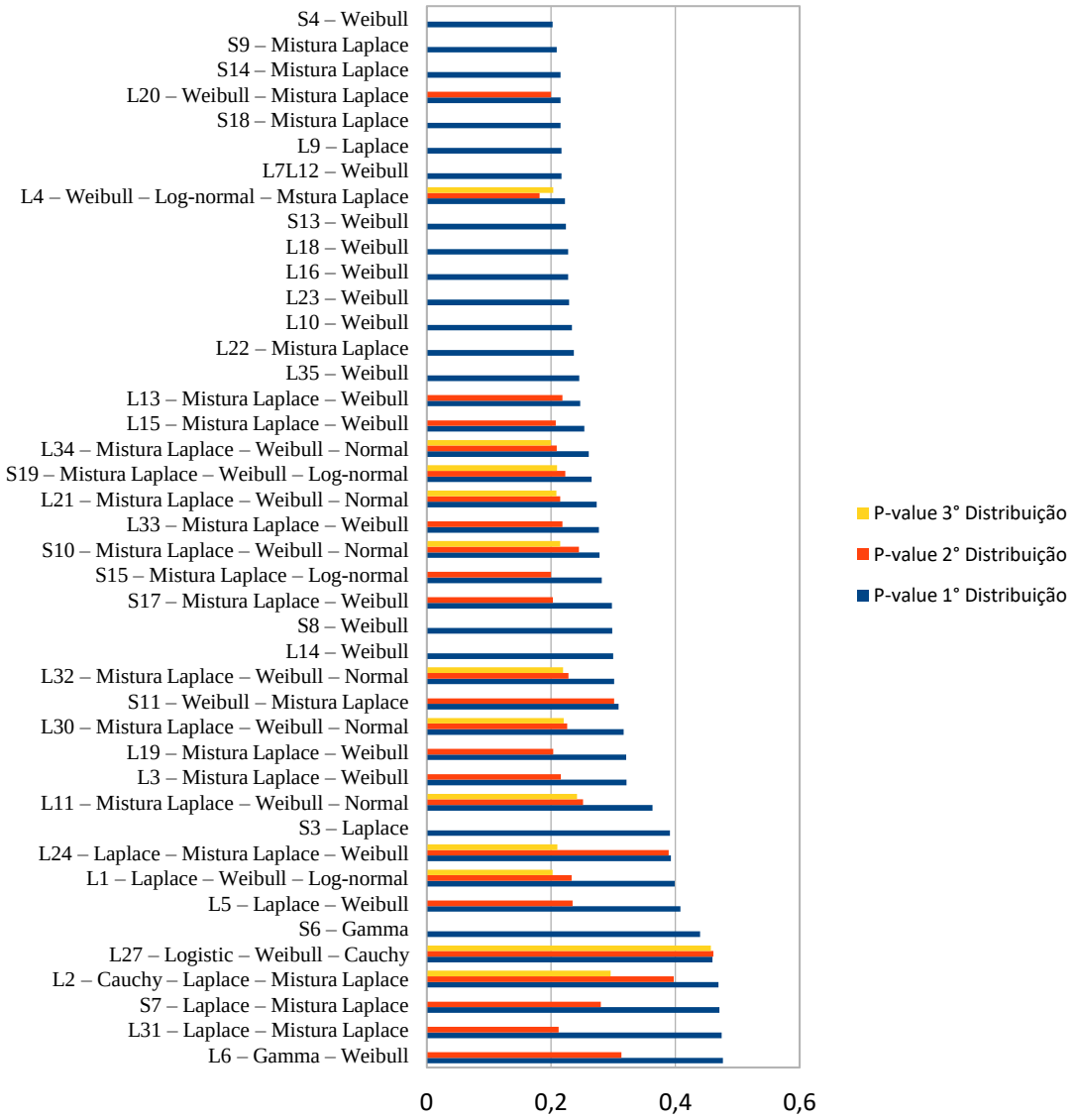
Fonte: O autor

4.1 SELEÇÃO DOS MODELOS

As seleções dos modelos foram baseadas nos valores *p* e selecionados somente os três melhores, que atingissem valores acime de 0,2. Algumas proteínas ficaram somente com um único modelo, isto ocorreu, pois, o ajuste da distribuição foi muito baixo. Foi identificado que as proteínas não seguem um padrão normal, ou seja, as proteínas seguem distribuições não paramétricas, esta condição também foi descrita em Parca et al., (2018). Mesmo para

distribuições com melhores ajustes selecionadas, o melhor resultado do valor *P* foi de 0,46. Quanto mais próximo de 1 for o valor *P* melhor os dados se comportam na distribuição selecionada (DEMORTIER, 2007). As distribuição selecionadas com base nos histogramas identificaram que seriam as melhores, confrontando com os dados dos testes estatísticos percebemos que o resultado é totalmente diferente como ilustrado na FIGURA 19.

FIGURA 19- Distribuições estatísticas selecionadas de acordo com o valor p e ranqueadas



Fonte: O autor

Foi necessário avaliar os modelos selecionados no gráfico ilustrado na FIGURA 19, para isso Han et al. (2018) relata o teste de log-likehood na avaliação de diversos modelos ilustrado no Quadro 2. Quanto menor o valor de log melhor os ajustes dos dados sobre a distribuição escolhida, para as distribuições Normal, Laplace, Gamma, Weibull, Logistic,

Cauchy e Log-normal, a partir dessa análise foi selecionada as proteínas L1, L2, L4, L5, L6, L9, L11, L21, L24, L27, L30, L31, L32, L34L S3, S6, S10, S15 e S19.

QUADRO 2- Modelos Estatísticos selecionados a partir do Score de log-likehood

Proteínas	Weibull	Laplace	Normal	Log-normal	Cauchy	Gamma
L1		-172597,2		-199374,9		
L2		-434038			-162340,2	
L4				-248710,7		
L5		-385331,6				
L6	-13857,03					-179914,5
L9		-372284,5				
L11	-15896,07		-20123,8			
L21	-15275,12		-200113,7			
L24	-14883,9	-445716,2				
L27	-13880,58				-175229,1	
L30	-13369,09		-181208,2			
L31		-411760,4				
L32	-14372,86		-183588,8			
L34	-11180,68		-1442292,3			
S3		-172671,5				
S6						-203528,9
S7		-453054				
S10	-13229,11		-174717,3			
S15				-182386,8		
S19	0			-189090,3		

Fonte: O autor

Confrontando os dois teste utilizados na identificação dos modelos estatísticos, observarmos que com o valor *P*, das distirbuições Cauchy para a proteína L2; Laplace para a proteínas L5, L31 e S7; Gamma para a proteína L6 e S6; Weibull para as proteínas L27 e S11; e por fim o modelo de mistura se comportou bem somente para a proteína S11, somente alguns modelos se ajustaram acima de 0.3 no valor P. Contudo quando avaliamos os modelos utilizando o *log-likhood*, a quantidade de distribuições selecionadas foram L1, L2, L4, L5, L6, L9, L11, L21, L24, L27, L30, L31, L32, L34, S3, S6, S10, S15 e S19 que se mostraram bem avaliados, isso ocorre pois a forma de medida entre os dois métodos e totalmente distinto e as distribuições utilizadas não mostram o comportamento ideal das proteínas.

4.2 MODELOS GERADOS A PARTIR DO ALGORITMO GENÉTICO

Os modelos gerados a partir do algoritmo genético, e das proteínas pré-identificadas geraram diversos classificadores na qual somente quem tivesse o AIC mais baixo foi selecionado como o melhor, validação já utilizados nos trabalhos de Aurélio Cavalcanti Pacheco (1999), Coelho (2013), Koza (1997), Zhou, Liu e Wong (2004).

Para a modelagem da primeira configuração do Algoritmo Genético, com a população de tamanho 20 e iteração fixado em 20 e 50, foi selecionado as proteínas L6, L7a,

L7, L12, L19, L11, L4, L15, L16, L18, L19, L20, L21, L23, L27, L28, L31, L32, L35, L36, S5, S6, S11, S13, S14, S15, S16, S17, S19, S21 e S22 para o treinamento. Na qual a matriz confusão do modelo gerado ilustrado na Tabela 4, mostra que a categoria VERDADEIRO – VERDADEIRO foi de 285247 classificações corretas índice bastante elevado, enquanto o índice FALSO – FALSO se mostrou baixo em relação ao anterior, no restante os índices se mostraram confuso entre as classificações.

TABELA 4- Matriz confusão da geração 20-20 e 20-50

PREDIÇÃO	FALSO	VERDADEIRO
FALSO	182122	80125
VERDADEIRO	152532	285247

Fonte: O autor

A acurácia para este modelo gerado foi de 67%, índice kappa de 32,79%, o que de acordo com Abraira (2001), De Ullibarri Galparsoro e Pita Fermández (1999) é um aprendizado insatisfatório para a classificação. A sensibilidade alcançada com o modelo foi de 0,54%, demonstrando que o modelo não é muito sensível aos dados, e a especificidade de 7,8%, demonstrou que o modelo errou mais que acertou, pois relatado por Faceli et al. (2011) corresponde aos falsos positivos. O AIC obtido foi de 860735 e o desvio foi de 969094 com 700025 graus de liberdade, e como podemos notar o desvio é muito superior indicando a existência de sobreposição. Então a função do modelo Binomial possui a seguinte estrutura:

- $Y \sim Bin(Y \mid N, P)$
- $Y = \beta_0 + \beta_1 tof_1$

Onde a variável *tof* representa as intensidades das proteínas L6, L7a, L7, L12, L19, L11, L4, L15, L16, L18, L19, L20, L21, L23, L27, L28, L31, L32, L35, L36, S5, S6, S11, S13, S14, S15, S16, S17, S19, S21 e S22, utilizadas no treinamento do modelo. Os resultados do ajuste no R estão descritos na Tabela 5.

TABELA 5 - - Estimativas dos parâmetros do modelo ajustado no R aos dados de proteínas ribossomais da base PUKYU

	Estimativa	Error Padrão	Valor-Z	Valor-P
Interceto	2,19E+00	7,50E-03	292	< 2e-16
tof	-1,45E-04	4,96E-07	-292,5	< 2e-16

Fonte: O autor

Para a geração dos modelos com a população de tamanho 40 e a iteração fixada em 20 a 50, selecionaram a seguintes proteínas L4, L7, L7a, L7ae, L11, L13, L19, L24, L32, L33, L35, S3, S4, S5, S12, S14, S21 e S22, ao todo 18 proteínas selecionadas. Ocorrendo muitos falsos positivos no modelo como ilustrado na matriz confusão da Tabela 6. A acurácia de 71%

mostrou-se bastante promissora, mas o índice kappa do modelo foi de -0,0023, que demonstra que o modelo é fraco (ABRAIRA, 2001; DE ULLIBARRI GALPARSORO; PITA FERNÁNDEZ, 1999). A especificidade do modelo de 0,0011% demonstra que errou pouco e a sensibilidade de 99% demonstra a taxa de acerto da classe positiva.

TABELA 6 - Matriz confusão da geração 40-20 e 40-50

PREDIÇÃO	FALSO	VERDADEIRO
FALSO	503222	195181
VERDADEIRO	1393	230

Fonte: O autor

Os resultados do ajuste no R estão descritos na Tabela 7. O AIC obtido foi de 816652 e o desvio foi de 829037 com 700025 graus de liberdade, e como podemos notar o desvio é muito superior indicando a existência de sobreposição. Então a função do modelo Binomial possui a seguinte estrutura:

- $Y \sim Bin(Y \mid N, P)$
- $Y = \beta_0 + \beta_1 tof_1$

Onde a variável *tof* representa as intensidades das proteínas L4, L7, L7a, L7ae, L11, L13, L19, L24, L32, L33, L35, S3, S4, S5, S12, S14, S21 e S22, utilizadas no treinamento do modelo. Os resultados do ajuste no R estão descritos na Tabela 7.

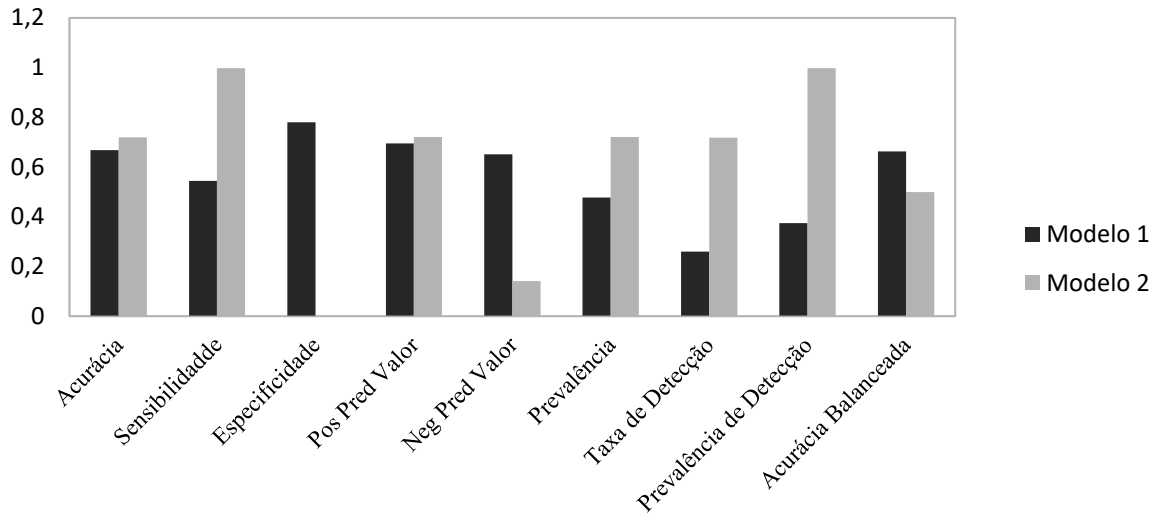
TABELA 7- Estimativas dos parâmetros do modelo ajustado no R aos dados de proteínas ribossomais da base PUKYU

	Estimativa	Error Padrão	Valor-Z	Valor-P
Interceto	-1,66E+00	7,06E-03	-235,1	< 2e-16
tof	4,78E-05	4,26E-07	111,4	< 2e-16

Fonte: O autor

Os dois modelos gerados mostraram sobreposição na classificação, contudo o primeiro modelo com 31 variáveis para construção do modelo revelou que a especificidade foi muito maior comparado com o segundo modelo de 18 variáveis, que foi próximo a zero. O balanceamento da acurácia do modelo 1 (um) foi muito maior, essa influência ocorreu devido a quantidade de variáveis para construção do modelo de predição mais informações são ilustradas na FIGURA 20.

FIGURA 20- Comparação entre os dois melhores modelos selecionados pelo AG.



Fonte: O autor

5 CONCLUSÃO

Neste estudo investigou-se as distribuições estatísticas das proteínas ribossomais. Os modelos de distribuições estatísticas mostraram-se em algumas proteínas bem ajustados levando em consideração os histogramas, mas o teste estatístico de CVM, não corroboravam com os histogramas. As melhores distribuições que se comportaram nos dados foram Cauchy para a proteína L2; Laplace para a proteínas L5, L31 e S7; Gamma para a proteína L6 e S6; Weibull para as proteínas L27 e S11; e o modelo de mistura se ajustou bem somente para a proteína S11.

A construção de um modelo linear generalizado, com o algoritmo genético para encontrar as melhores combinações das proteínas para o treinamento do classificador, se mostrou favorável, pois foram selecionados dois modelos classificatórios. Isto apresentou uma validade representativa de combinações para o classificador, diferenciando na acurácia e na especificidade dos modelos e os dois modelos obtiveram alto índice de sobreposição.

5.1 TRABALHOS FUTUROS

Os conhecimentos obtidos através do desenvolvimento desta dissertação podem ser consideravelmente ampliados através da utilização da mesma base de dados ou uma base de dados mais completa, com outros algoritmos de otimização específicos para identificação do comportamento dos dados. A disponibilidade de novas técnicas viabiliza novas metodologias de que podem reduzir o custo computacional na criação de novos modelos mais precisão e ideais para a aplicação.

Eis as sugestões para trabalhos futuros:

Testar outras distribuições estatísticas do tipo contínuo para verificar se alguma distribuição se ajusta melhor aos dados, bem como o teste de outros modelos misturas e modelos truncados.

Utilizar algoritmos bioinspirados para estimar os parâmetros e estimar os modelos das distribuições dos dados, para isso é necessário a implementação do algoritmo EDA – Estimation of Distribution Algorithms, este algoritmo é do tipo evolucionário que combina a aprendizado de máquina e estatística para estimação.

REFERÊNCIAS

ABRAIRA, V. El índice kappa. **Semergen-Medicina de Familia**, v. 27, n. 5, p. 247–249, 2001.

AHSANULLAH, M. **Characterizations of Univariate Continuous Distributions**. 1. ed. [s.l.] Atlantis Press, 2017.

ALMEIDA, O. G. G.; DE MARTINIS, E. C. P. Bioinformatics tools to assess metagenomic data for applied microbiology. **Applied Microbiology and Biotechnology**, v. 103, n. 1, p. 69–82, 25 jan. 2019.

AURÉLIO CAVALCANTI PACHECO, M. Algoritmos genéticos: principios e aplicações. **ICA: Laboratório de Inteligência Computacional Aplicada. Departamento de Engenharia Elétrica. Pontificia Universidade Católica do Rio de Janeiro. Fonte desconhecida**, p. 28, 1999.

BALDI, P.; BRUNAK, S.; BACH, F. **Bioinformatics: the machine learning approach**. [s.l.] MIT press, 2001.

BARDGETT, R. D.; DE VRIES, F. T.; VAN DER PUTTEN, W. H. Soil Biodiversity and Ecosystem Functioning. In: **Microbial Biomass**. [s.l.] WORLD SCIENTIFIC (EUROPE), 2017. p. 119–140.

BERG, J. M.; TYMOCZKO, J. L.; STRYER, L. **Bioquímica**. 7. ed. [s.l: s.n.]. v. 7

BRYC, W. **The normal distribution: characterizations with applications**. [s.l.] Springer Science & Business Media, 2012. v. 100

CHAMIZO, S. et al. Soil inoculation with cyanobacteria: reviewing its' potential for agriculture sustainability in Drylands. **Agri. Res. Tech Open Access. J**, v. 18, p. 556046, 2018.

CHATFIELD, C. **Statistics for technology: a course in applied statistics**. [s.l.] Routledge, 2018.

COELHO, G. V. V. **Seleção de características usando algoritmos genéticos para classificação de imagens de textos em manuscritos e impressos**. [s.l.] Universidade Federal de Pernambuco, 2013.

CORDEIRO, G. M.; DEMÉTRIO, C. G. B. Modelos lineares generalizados e extensões. **Sao Paulo**, v. 33, 2008.

COX, V. Exploratory data analysis. In: **Translating Statistics to Make Decisions**. [s.l.] Springer, 2017. p. 47–74.

CROW.; SHIMIZU, K. **Lognormal Distributions: Theory and Applications**. [s.l.] Routledge, 2018.

DE BRUYNE, K. et al. Bacterial species identification from MALDI-TOF mass spectra through data analysis and machine learning. **Systematic and Applied Microbiology**, v. 34, n. 1, p. 20–29, fev. 2011.

DE ULLIBARRI GALPARSORO, L.; PITA FERNÁNDEZ, S. Medidas de concordancia: el Índice de Kappa. **Cad Aten Primaria**, v. 6, p. 169–171, 1999.

DEB, K. **Multi-objective optimization using evolutionary algorithms**. [s.l.] John Wiley & Sons, 2001. v. 16

DEMORTIER, L. **P Values: what they are and how to use them**. [s.l: s.n.].

DOBSON, A. J.; BARNETT, A. G. **An introduction to generalized linear models**. [s.l.] Chapman and Hall/CRC, 2008.

DONG, H. et al. A novel hybrid genetic algorithm with granular information for feature selection and optimization. **Applied Soft Computing**, v. 65, p. 33–46, 2018.

EVERITT, B. S. Finite Mixture Distributions. In: **Wiley StatsRef: Statistics Reference Online**. Chichester, UK: John Wiley & Sons, Ltd, 2014. v. 66p. 178.

FACELI, K. et al. Inteligência Artificial: Uma abordagem de aprendizado de máquina. 2011.

FAMOYE, F. **Continuous Univariate Distributions, Volume 1**Taylor & Francis Group, , 1995.

FORBES, C. et al. **Statistical distributions**. [s.l.] John Wiley & Sons, 2011.

GALVÃO, C. W. et al. **PUKYU - Banco de dados de Massa Molecular de Proteínas Ribossomais**Ponta Grossa, 2017. Disponível em: <https://gru.inpi.gov.br/pePI/servlet/ProgramaServletController?Action=detail&CodPedido=23657&SearchParameter=PUKYU>

GANS, J. Computational Improvements Reveal Great Bacterial Diversity and High Metal Toxicity in Soil. **Science**, v. 309, n. 5739, p. 1387–1390, ago. 2005.

GEN, M.; LIN, L. Genetic Algorithms. **Wiley Encyclopedia of Computer Science and Engineering**, p. 1–15, 2007.

GOLDBARG, E.; GOLDBARG, M.; LUNA, H. **Otimização combinatória e meta-heurísticas: algoritmos e apliacações**. [s.l.] Elsevier Brasil, 2017.

GOULART, V. A. M.; RESENDE, R. R. MALDI-TOF : uma ferramenta revolucionária para as análises clínicas e pesquisa do câncer. p. 1–5, 2013.

GREEN, D.; ALETI, A.; GARCIA, J. The nature of nature: why nature-inspired algorithms work. In: **Nature-Inspired Computing and Optimization**. [s.l.] Springer, 2017. p. 1–27.

GUPTA, R. D.; KUNDU, D. Generalized logistic distributions. **Journal of Applied Statistical Science**, v. 18, n. 1, p. 51, 2010.

HAN, Z.-Y. et al. Unsupervised generative modeling using matrix product states. **Physical Review X**, v. 8, n. 3, p. 31012, 2018.

HAWLEY, A. K. et al. Diverse Marinimicrobia bacteria may mediate coupled biogeochemical cycles along eco-thermodynamic gradients. **Nature Communications**, v. 8, n. 1, p. 1507, dez. 2017.

KOZA, J. R. Genetic programming. 1997.

KRAMER, O. **Genetic algorithm essentials**. [s.l.] Springer, 2017. v. 679

KRISHNAMOORTHY, K. **Handbook of statistical distributions with applications**. Second edi ed. [s.l.] CRC Press, 2016.

KUMAR, S.; JAIN, S.; SHARMA, H. Genetic algorithms. **Advances in Swarm Intelligence for Optimizing Problems in Computer Science**, p. 27–52, 2018.

LEONARD KAUFMAN, P. J. R. **Finding Groups in Data: An Introduction to Cluster Analysis**. 1. ed. [s.l.] Wiley-Interscience, 2005.

LI, Q.; MACCOSS, M. J.; STEPHENS, M. A nested mixture model for protein identification using mass spectrometry. **Annals of Applied Statistics**, v. 4, n. 2, p. 962–987, 2010.

LIYANAGE, R. et al. Matrix-assisted ionization Fourier transform mass spectrometry for the analysis of lipids. **Rapid Communications in Mass Spectrometry**, 2019.

LLOYD JOHNSON, N.; KOTS, S.; BALAKRISHNAN, N. **Continuous Univariate Distributions, Vol. 2**. Volume 2 ed. [s.l.] Wiley-Interscience, 1995.

LOPER, J. E.; SUSLOW, T. V; SCHROTH, M. N. Lognormal distribution of bacterial populations in the rhizosphere. **Phytopathology**, v. 74, n. 12, p. 1454–1460, 1984.

LÓPEZ-FERNÁNDEZ, H. et al. Mass-Up: an all-in-one open software application for MALDI-TOF mass spectrometry knowledge discovery. **BMC bioinformatics**, v. 16, n. 1, p. 318, 2015.

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

LU, D.; MORAN, E.; BATISTELLA, M. Linear mixture model applied to Amazonian vegetation classification. **Remote Sensing of Environment**, v. 87, n. 4, p. 456–469, 2003.

MANOGARAN, G. et al. Machine learning based big data processing framework for cancer diagnosis using hidden Markov model and GM clustering. **Wireless personal communications**, v. 102, n. 3, p. 2099–2116, 2018.

MCLACHLAN, G. J.; BASFORD, K. E. **Mixture models: Inference and applications to clustering**. [s.l.] Marcel Dekker, 1988. v. 84

MIYAZAWA, F. K.; DE SOUZA, C. C. **Introdução a Otimização Combinatória**. Jornadas de Atualizaçao em Informática-Congresso da Sociedade Brasileira de Computação JAI-SBC. **Anais**...2015

NG, W. **Conserved mass peaks in MALDI-TOF mass spectra of bacterial species at the genus and species levels**. [s.l: s.n.]. Disponível em: <https://peerj.com/preprints/3524/>.

NOGUEIRA LONDE, L.; SALES DIAS, L.; SOARES DA SILVA, A. Avaliação de impacto ambiental aplicado às atividades agrícolas na lagoa grande no município de Janaúba - MG. **Revista Desenvolvimento Social No**, v. 1401, p. 2179–6807, 2015.

OLIVIER CHAPELLE BERNHARD SCHÖLKOPF, A. Z. **Semi-Supervised Learning**. [s.l.] MIT Press, 2006.

PANDEY, A. et al. 16S rRNA gene sequencing and MALDI-TOF mass spectrometry based comparative assessment and bioprospection of psychrotolerant bacteria isolated from high altitudes under mountain ecosystem. **SN Applied Sciences**, v. 1, n. 3, p. 278, 2019.

PARCA, L. et al. Quantifying compartment-associated variations of protein abundance in proteomics data. **Molecular systems biology**, v. 14, n. 7, 2018.

PSAROULAKI, A.; CHOCHLAKIS, D. Use of MALDI-TOF mass spectrometry in the battle against bacterial infectious diseases: recent achievements and future perspectives. **Expert Review of Proteomics**, v. 15, n. 7, p. 537–539, 3 jul. 2018.

RINNE, H. **The Weibull distribution: a handbook**. [s.l.] Chapman and Hall/CRC, 2008.

RINNE, H. **The Weibull Distribution: A Handbook (2009)**. 1. ed. [s.l: s.n.].

RODRIGUES, C.; PASSET, V.; BRISSE, S. Identification of Klebsiella pneumoniae complex members using MALDI-TOF mass spectrometry. **bioRxiv**, p. 350579, 2018.

RODRIGUEZ, G. **Generalized Linear Models**. [s.l: s.n.].

ROSS, S. M. **Introduction to probability and statistics for engineers and scientists**. [s.l.] Academic Press, 2014.

SAMPEDRO, A.; CEBALLOS MENDIOLA, J.; ALIAGA MARTÍNEZ, L. **MALDI-TOF Commercial Platforms for Bacterial Identification**. [s.l.] Elsevier Inc., 2018.

SAMUEL KOTZ TOMAZ J. KOZUBOWSKI, K. P. (AUTH. . **The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance**. 1. ed. [s.l.] Birkhäuser Basel, 2001.

SANTOS, F. DOS; DOS SANTOS, F. **Algoritmo KNN na imputação de dados de espectros de Massa do Tipo MALDI-TOF**. [s.l: s.n.].

SANTOS, K. D.; MARCELOS, W. R. PROBLEMAS DE OTIMIZAÇÃO (OU PROBLEMAS DE MAXIMIZAÇÃO E MINIMIZAÇÃO). **Anais do EVINCI-UniBrasil**, v. 1, n. 3, p. 320, 2016.

SCHLOSS, P. D.; HANDELSMAN, J. Toward a census of bacteria in soil. **PLoS Computational Biology**, v. 2, n. 7, p. 786–793, jul. 2006.

SCRUCCA, L. GA: a package for genetic algorithms in R. **Journal of Statistical Software**, v. 53, n. 4, p. 1–37, 2013.

SCRUCCA, L. Identifying connected components in Gaussian finite mixture models for clustering. **Computational Statistics & Data Analysis**, v. 93, n. xxxx, p. 5–17, 2016.

SHAMEER, S.; PRASAD, T. Plant growth promoting rhizobacteria for sustainable agricultural practices with special reference to biotic and abiotic stresses. **Plant growth regulation**, v. 84, n. 3, p. 603–615, 2018.

SHEN, C. et al. A hierarchical statistical model to assess the confidence of peptides and proteins inferred from tandem mass spectrometry. **Bioinformatics**, v. 24, n. 2, p. 202–208, jan. 2008.

SILVA, L. C. Aplicação de espectrometria de massa no controle de qualidade de inoculantes comerciais agrícolas contendo bactérias promotoras de crescimento vegetal. 2014.

SIVANANDAM, S. N.; DEEPA, S. N. **Introduction to Genetic Algorithms**. 1. ed. [s.l.] Springer, 2007.

SIVANANDAM, S. N.; DEEPA, S. N. Genetic algorithms. In: **Introduction to genetic algorithms**. [s.l.] Springer, 2008. p. 15–37.

SOUZA, L. B. DE. **Análise da calogênese de paricá por modelos de regressão logistica**. [s.l.] Universidade Federal de Alfenas, 2015.

SYSWERDA, G. Simulated crossover in genetic algorithms. In: **Foundations of genetic algorithms**. [s.l.] Elsevier, 1993. v. 2p. 239–255.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introdução ao datamining: mineração de dados**. [s.l.] Ciência Moderna, 2009.

THAJUDDIN, N. **Bacterial morphology**.

TOMACHEWSKI, D. **Utilização de aprendizagem de máquina para classificação de bactérias através de proteínas Ribossomais**. [s.l.] Universidade Estadual de Ponta Grossa, 2017.

TOMACHEWSKI, D. et al. Ribopeaks: a web tool for bacterial classification through m/z data from ribosomal proteins. **Bioinformatics (Oxford, England)**, v. 34, n. 17, p. 3058–3060, 1 set. 2018.

UMBARKAR, A. J.; SHETH, P. D. CROSSOVER OPERATORS IN GENETIC ALGORITHMS: A REVIEW. **ICTACT journal on soft computing**, v. 6, n. 1, 2015.

VILLACORTA, P. J. et al. Cluster-based comparison of the peptide mass fingerprint obtained by MALDI-TOF mass spectrometry. A case study: long-term stability of rituximab. **Analyst**, v. 140, n. 5, p. 1717–1730, 2015.

WANG, T. et al. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. **BMC Bioinformatics**, v. 20, n. 1, p. 1–16, 2019.

WOESE, C. R.; KANDLER, O.; WHEELIS, M. L. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. **Proceedings of the National Academy of Sciences**, v. 87, n. 12, p. 4576–4579, 1990.

WOLK, D. M.; CLARK, A. E. Matrix assisted laser desorption time of flight mass spectrometry. **Clinics in Laboratory Medicine**, v. 38, n. 3, p. 471–486, 2018.

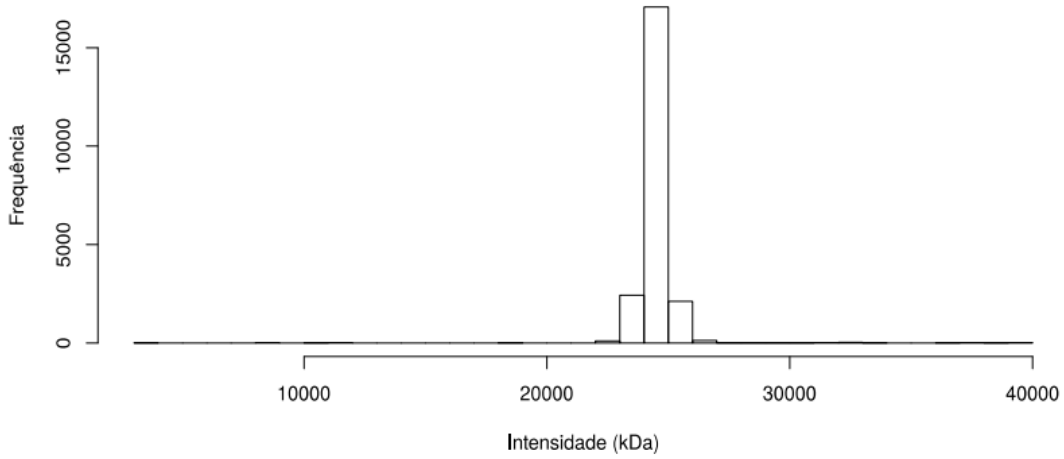
YADAV, K. K.; SARKAR, S. Biofertilizers, impact on soil fertility and crop productivity under sustainable agriculture. **Environment and Ecology**, v. 37, n. 1, p. 89–93, 2019.

ZHOU, X.; LIU, K.-Y.; WONG, S. T. C. Cancer classification and prediction using logistic regression with Bayesian gene selection. **Journal of Biomedical Informatics**, v. 37, n. 4, p. 249–259, 2004.

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

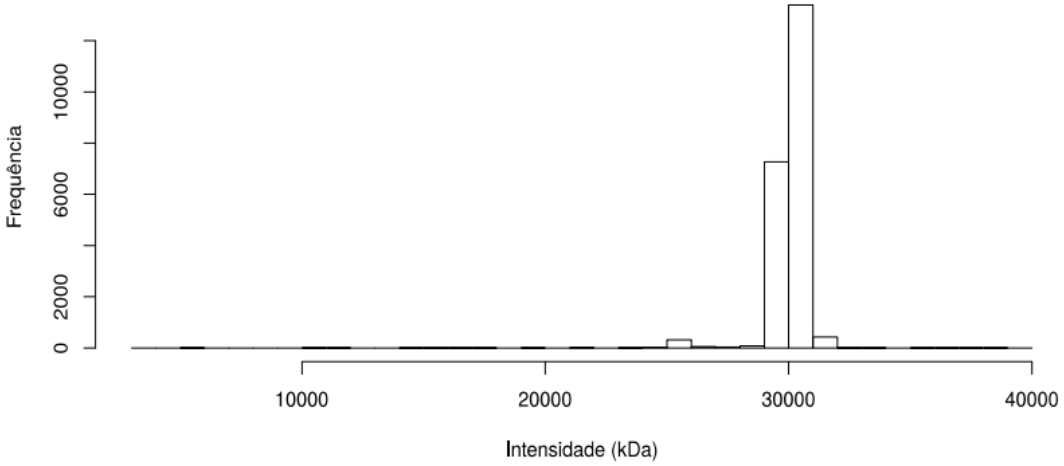
APÊNDICE A – HISTOGRAMA DAS PROTEÍNAS RIBOSSOMAIS

FIGURA 21 - Histograma Proteína Ribossomal L1



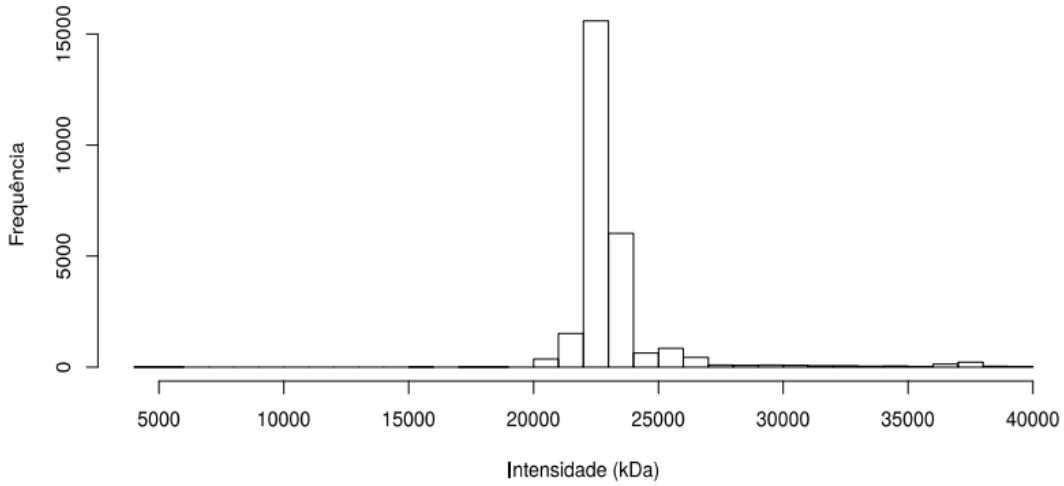
Fonte: O autor

FIGURA 22- Histograma Proteína Ribossomal L2



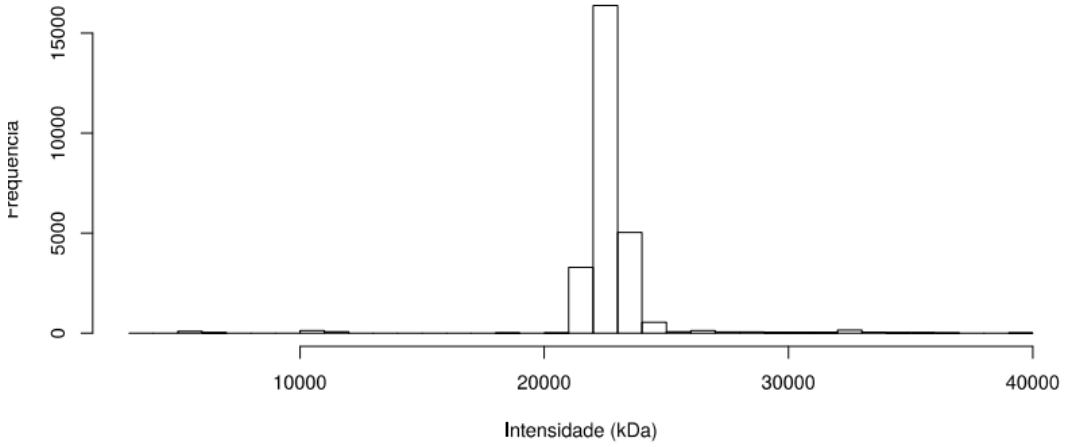
Fonte: O autor

FIGURA 23- Histograma Proteína Ribossomal L3



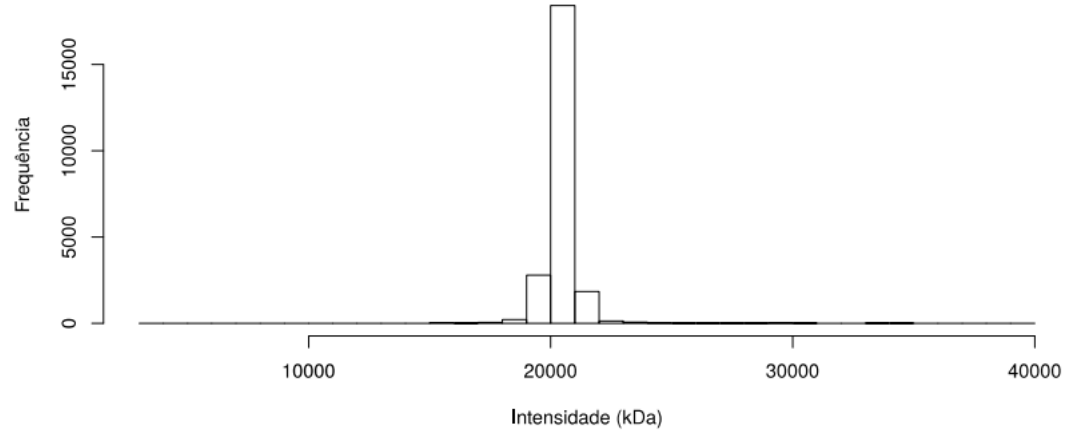
Fonte: O autor

FIGURA 24- Histograma Proteína Ribossomal L4



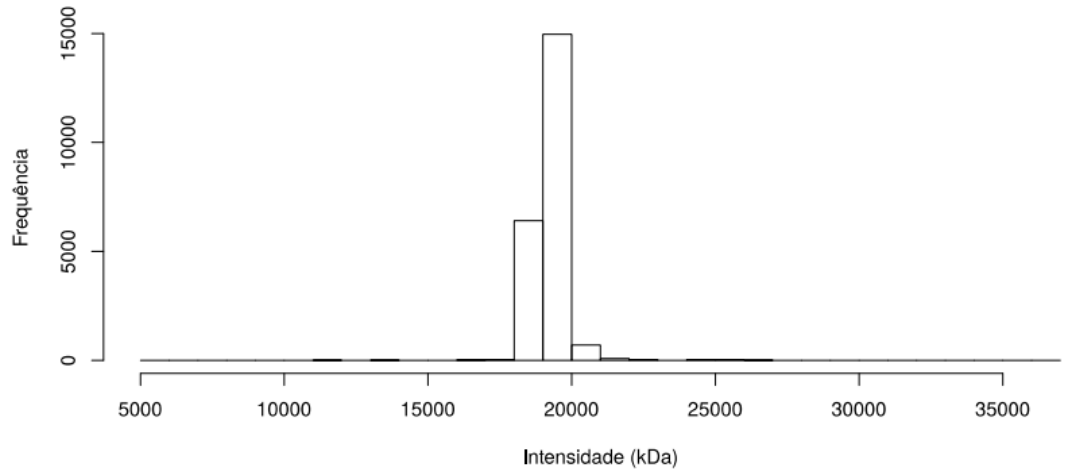
Fonte: O autor

FIGURA 25- Histograma Proteína Ribossomal L5



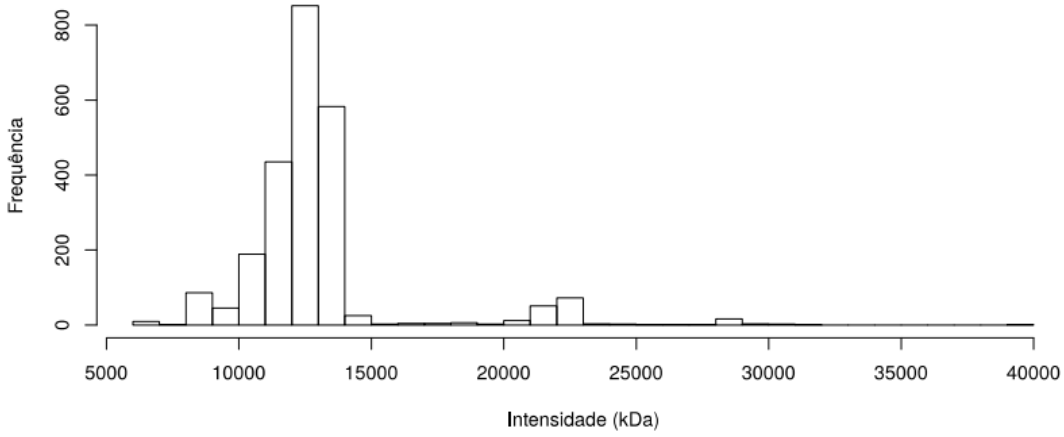
Fonte: O autor

FIGURA 26 - Histograma Proteína Ribossomal L6



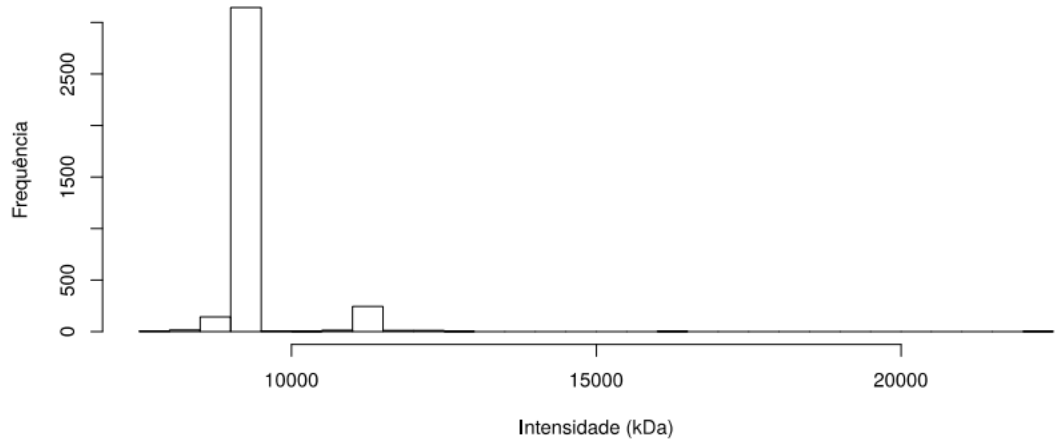
Fonte: O autor

FIGURA 27 - Histograma Proteína Ribossomal L7



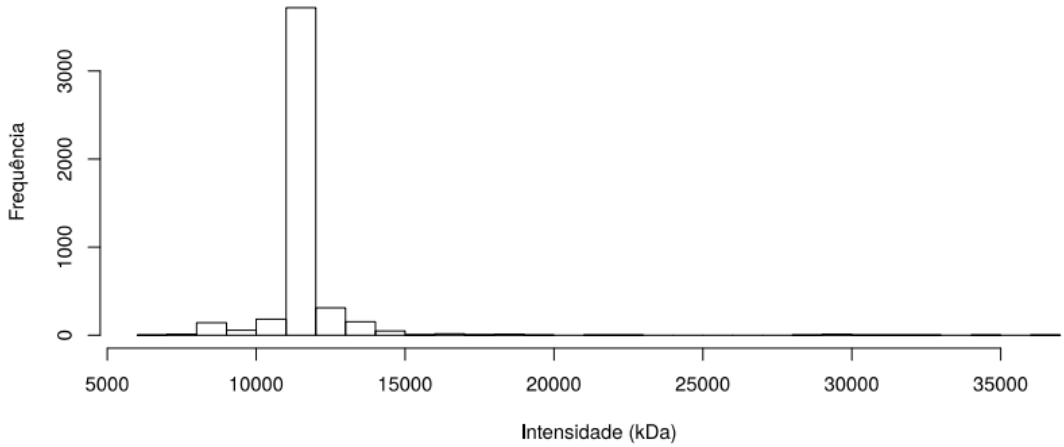
Fonte: O autor

FIGURA 28- Histograma Proteína Ribossomal L7a



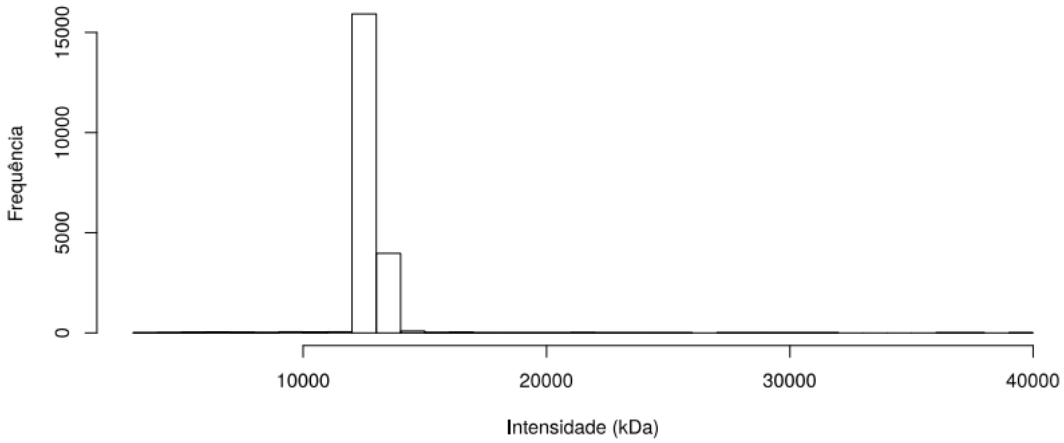
Fonte: O autor

FIGURA 29- Histograma Proteína Ribossomal L7ae



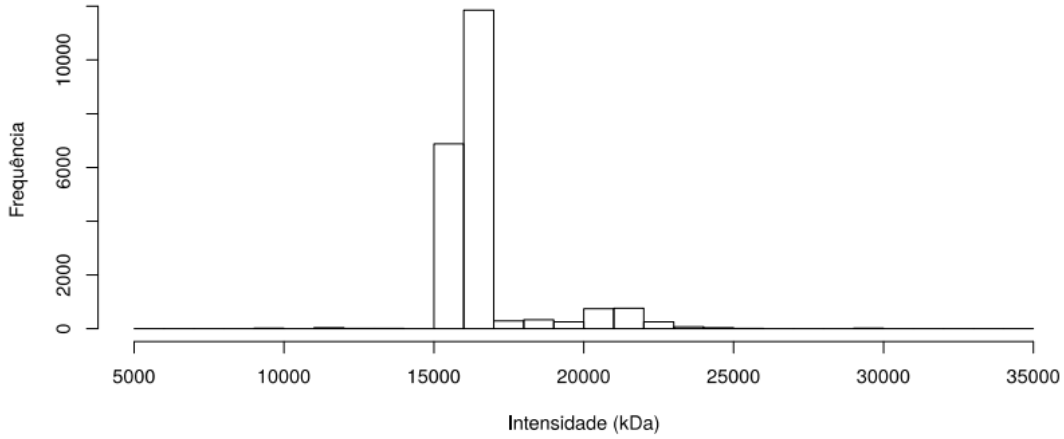
Fonte: O autor

FIGURA 30- Histograma Proteína Ribossomal L7L12



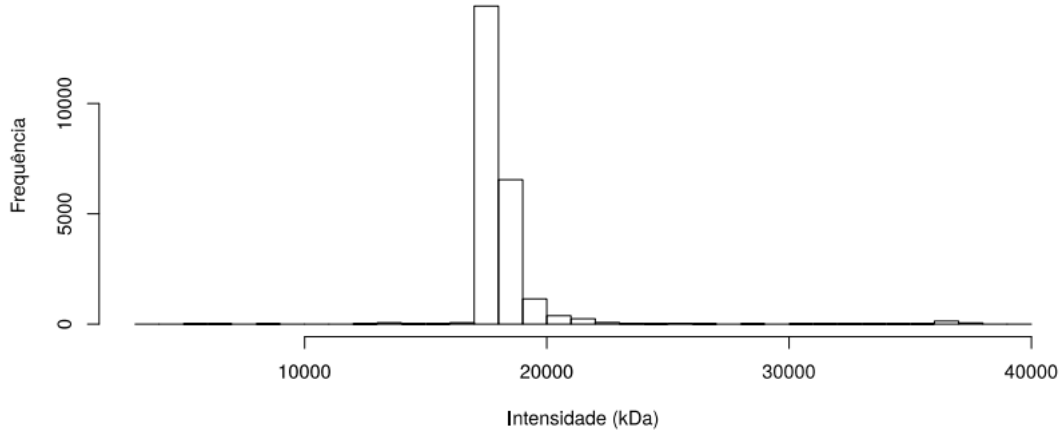
Fonte: O autor

FIGURA 31- Histograma Proteína Ribossomal L9



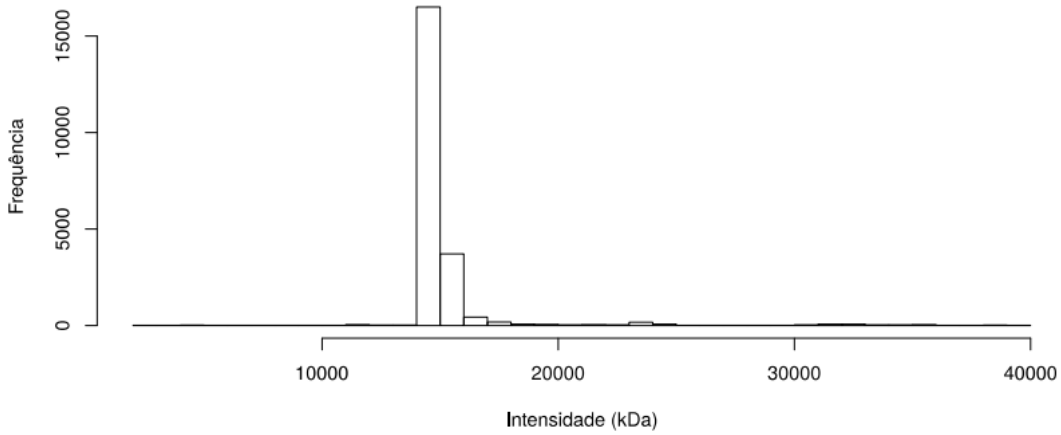
Fonte: O autor

FIGURA 32- Histograma Proteína Ribossomal L10



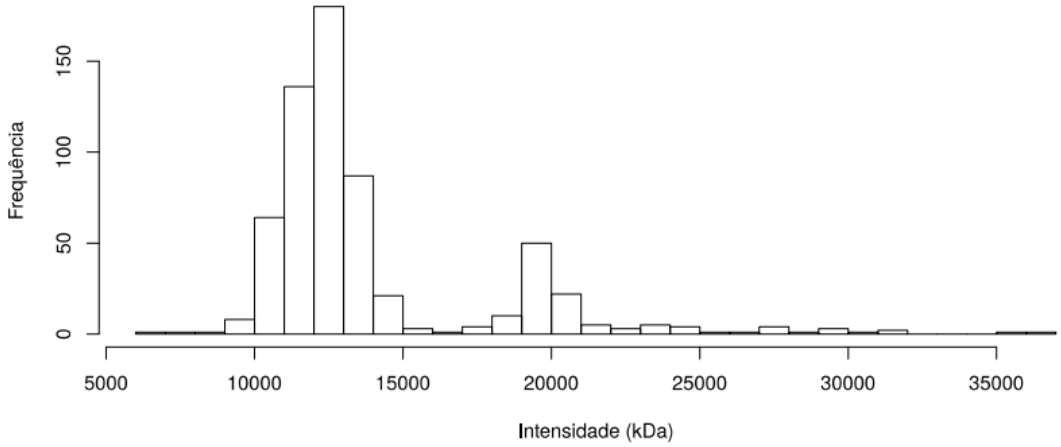
Fonte: O autor

FIGURA 33- Histograma Proteína Ribossomal L11



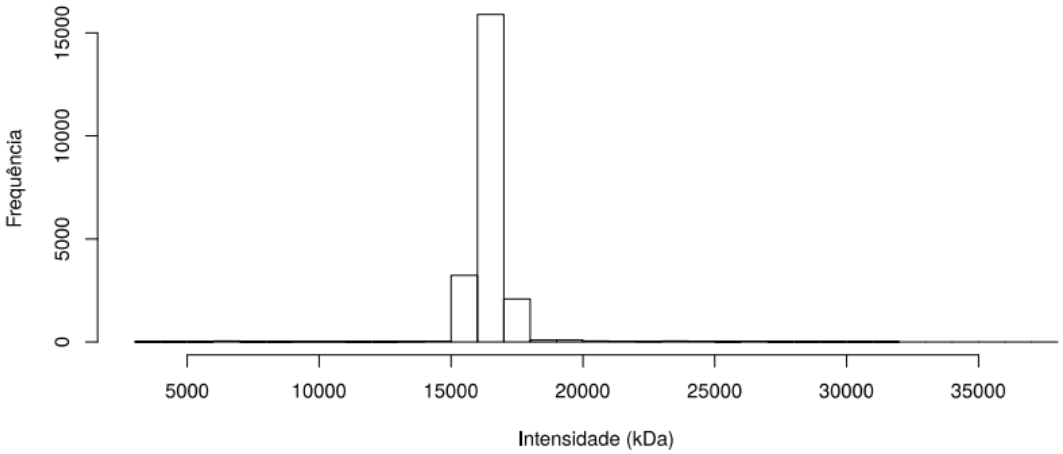
Fonte: O autor

FIGURA 34- Histograma Proteína Ribossomal L12



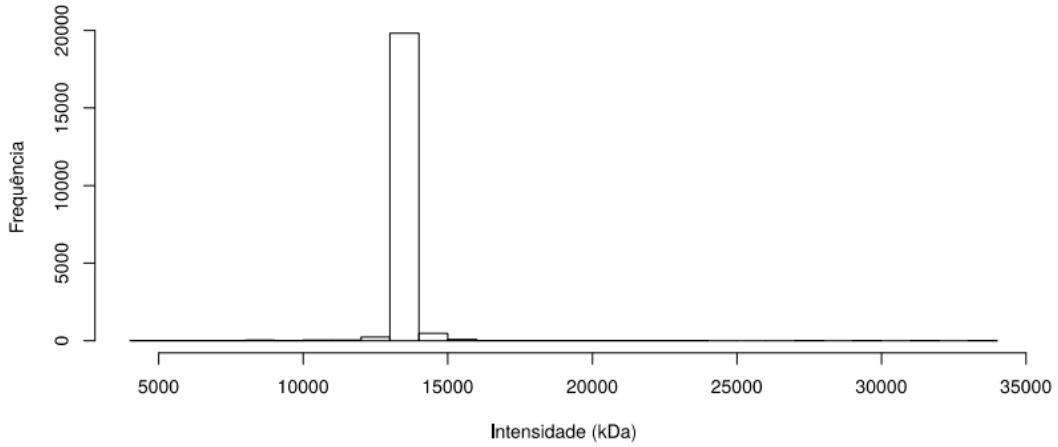
Fonte: O autor

FIGURA 35 - Histograma Proteína Ribossomal L13



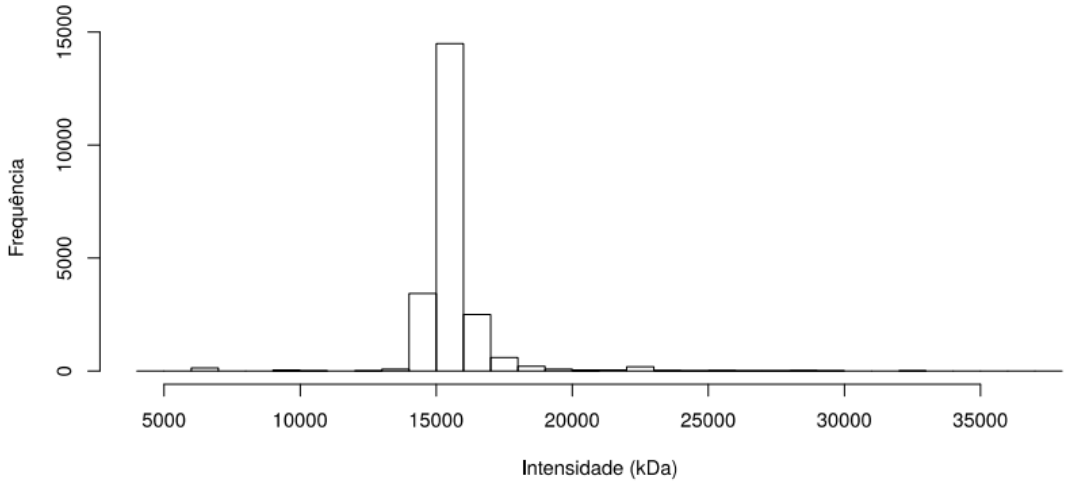
Fonte: O autor

FIGURA 36 - Histograma Proteína Ribossomal L14



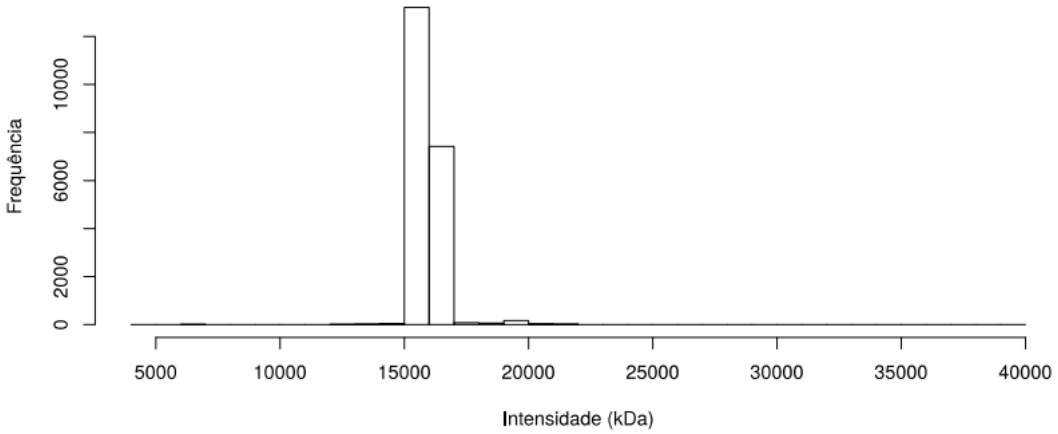
Fonte: O autor

FIGURA 37 - Histograma Proteína Ribossomal L15



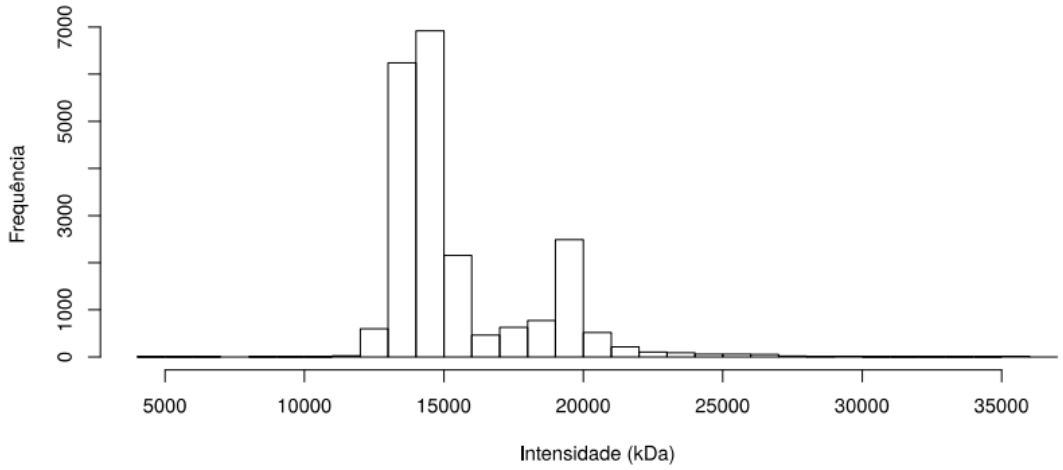
Fonte: O autor

FIGURA 38 - - Histograma Proteína Ribossomal L16



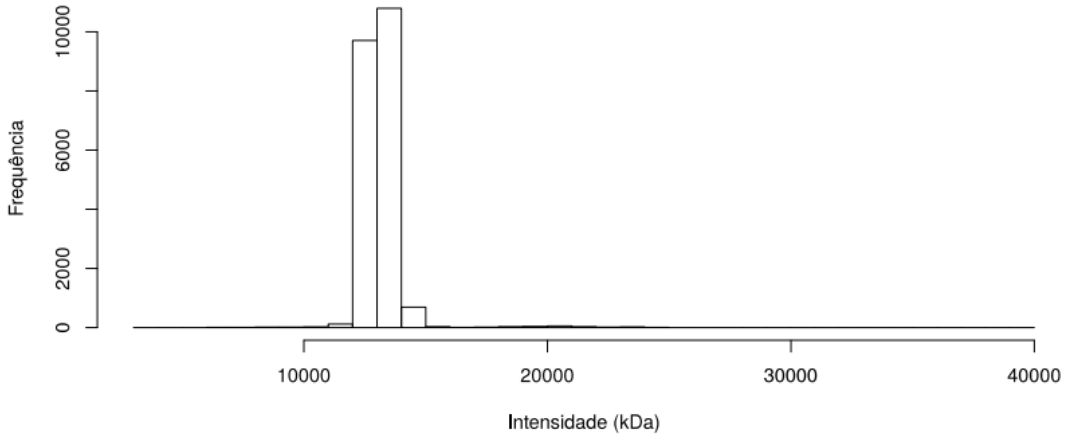
Fonte: O autor

FIGURA 39 - Histograma Proteína Ribossomal L17



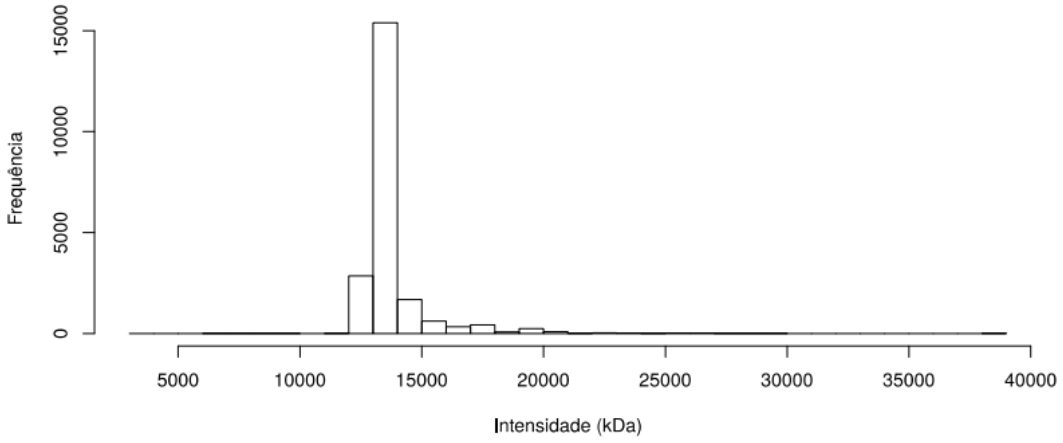
Fonte: O autor

FIGURA 40 - Histograma Proteína Ribossomal L18



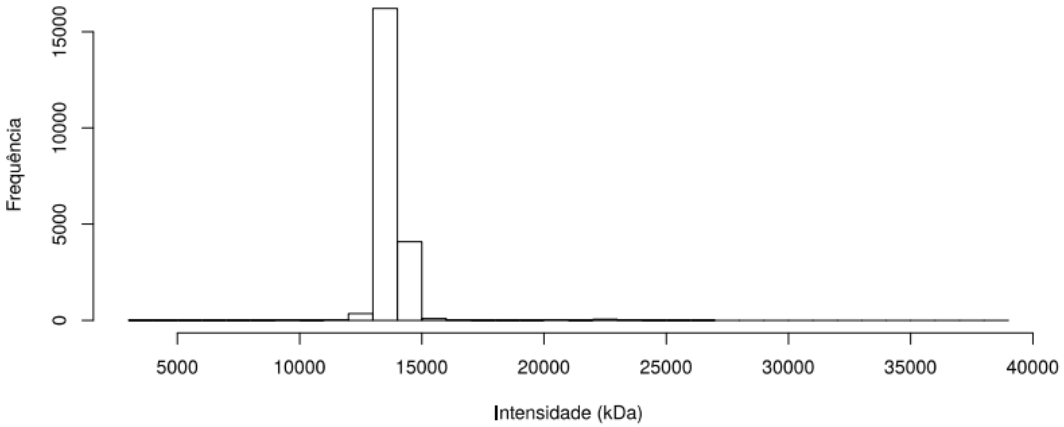
Fonte: O autor

FIGURA 41 - Histograma Proteína Ribossomal L19



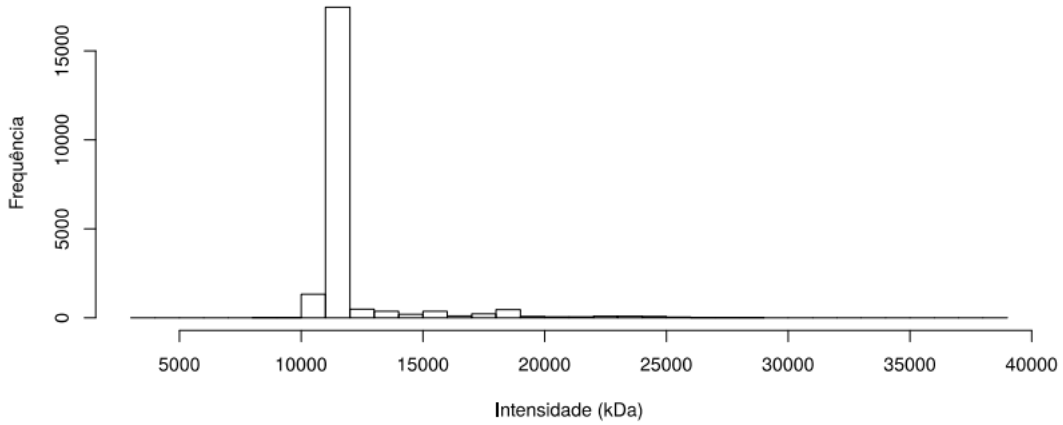
Fonte: O autor

FIGURA 42 - Histograma Proteína Ribossomal L20



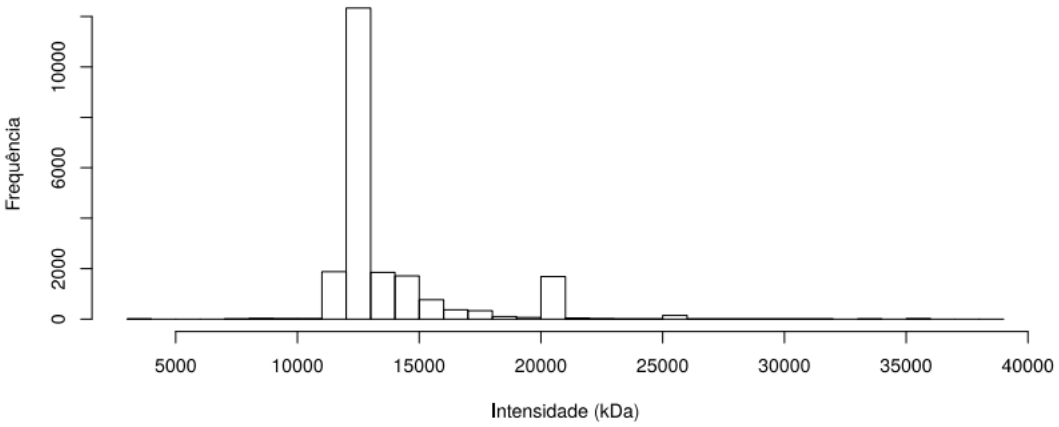
Fonte: O autor

FIGURA 43 - Histograma Proteína Ribossomal L21



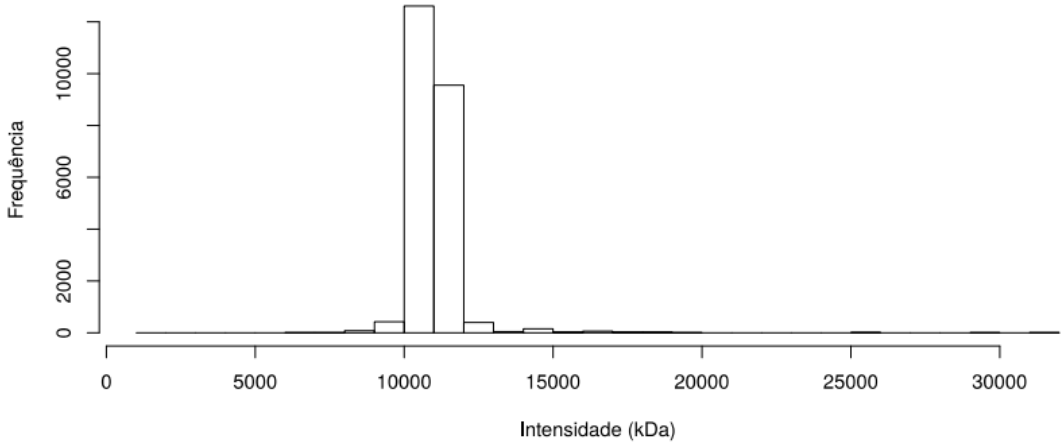
Fonte: O autor

FIGURA 44 - Histograma Proteína Ribossomal L22



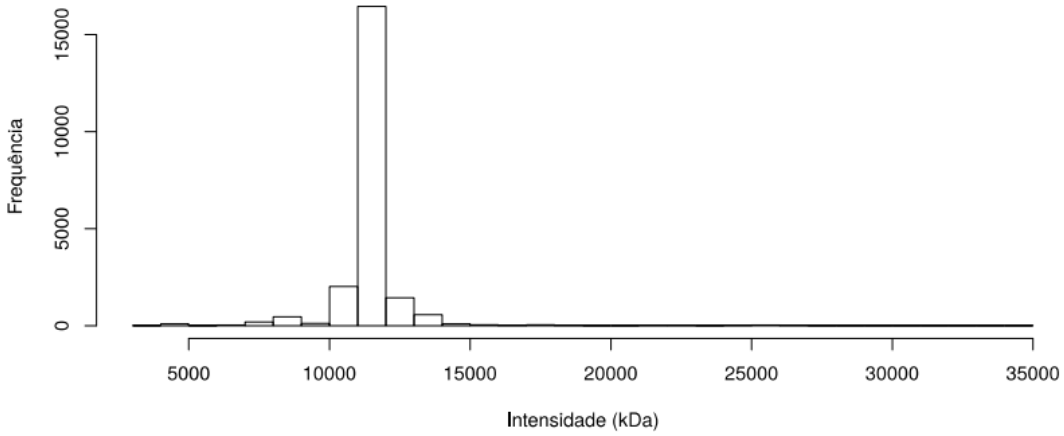
Fonte: O autor

FIGURA 45 - Histograma Proteína Ribossomal L23



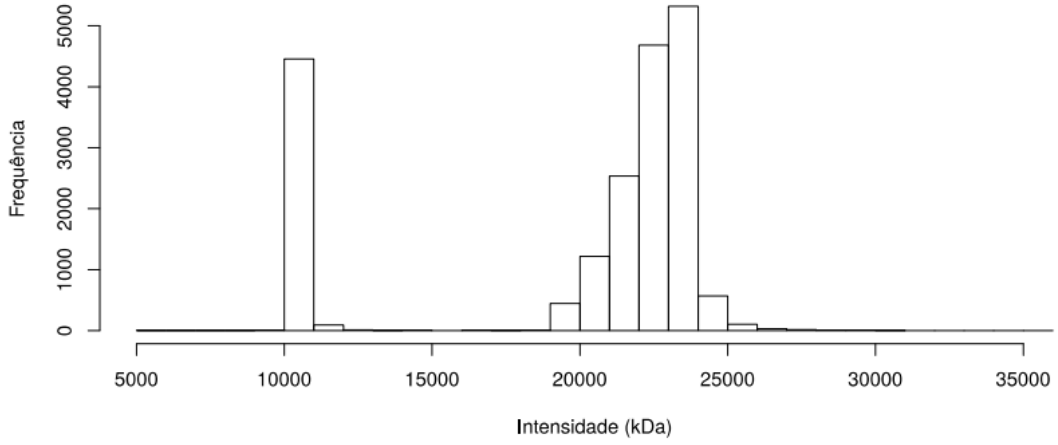
Fonte: O autor

FIGURA 46 - Histograma Proteína Ribossomal L24



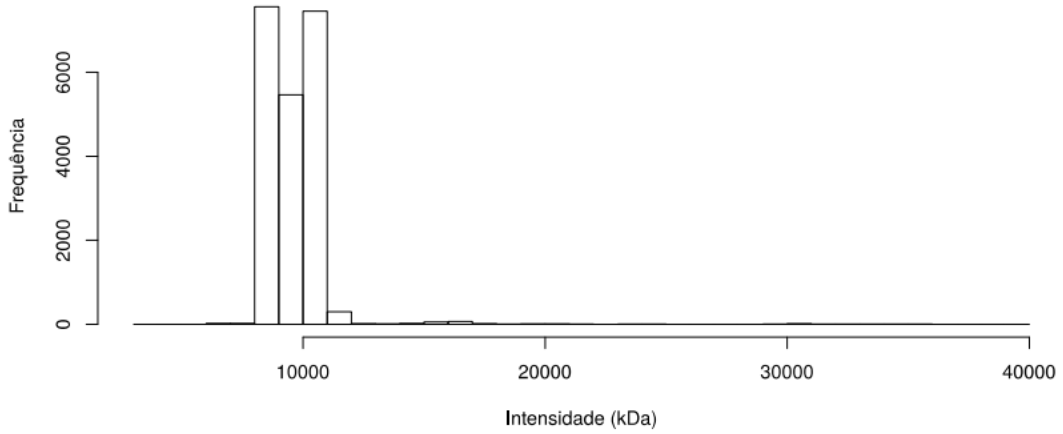
Fonte: O autor

FIGURA 47 - Histograma Proteína Ribossomal L25



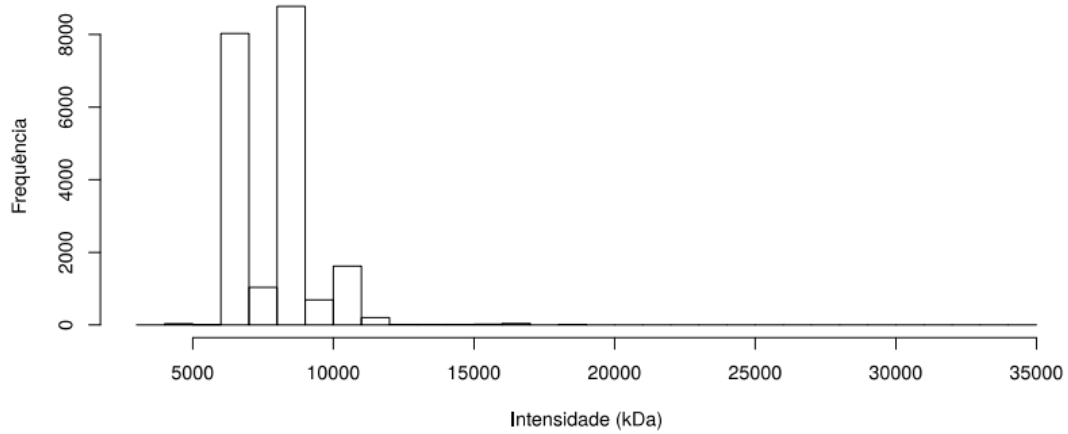
Fonte: O autor

FIGURA 48 - Histograma Proteína Ribossomal L27



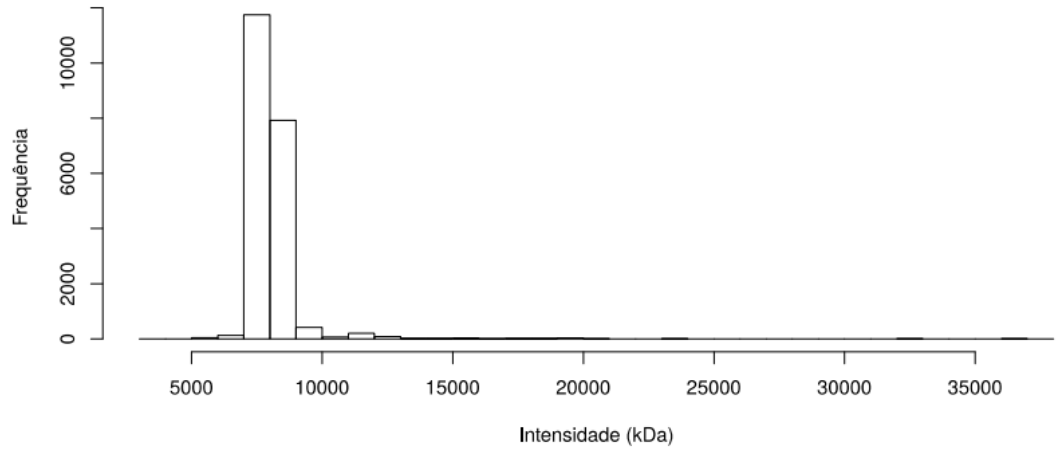
Fonte: O autor

FIGURA 49 - Histograma Proteína Ribossomal L28



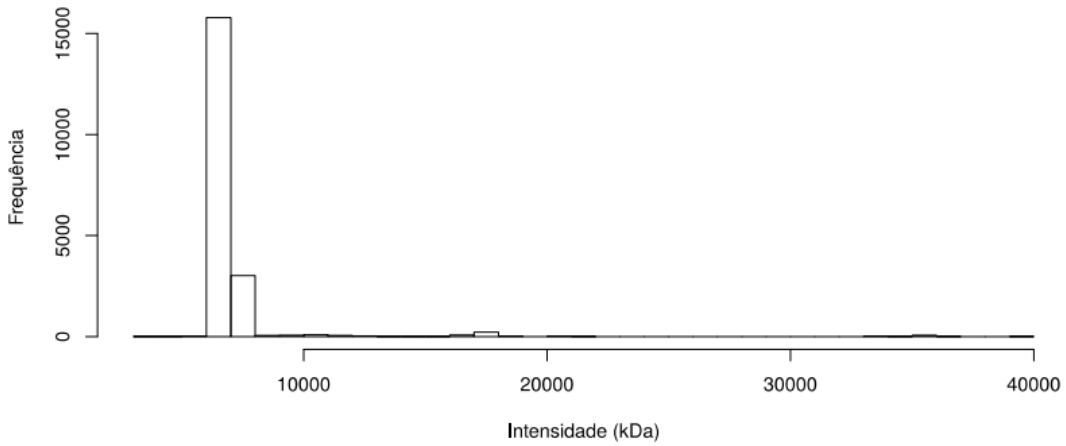
Fonte: O autor

FIGURA 50 - Histograma Proteína Ribossomal L29



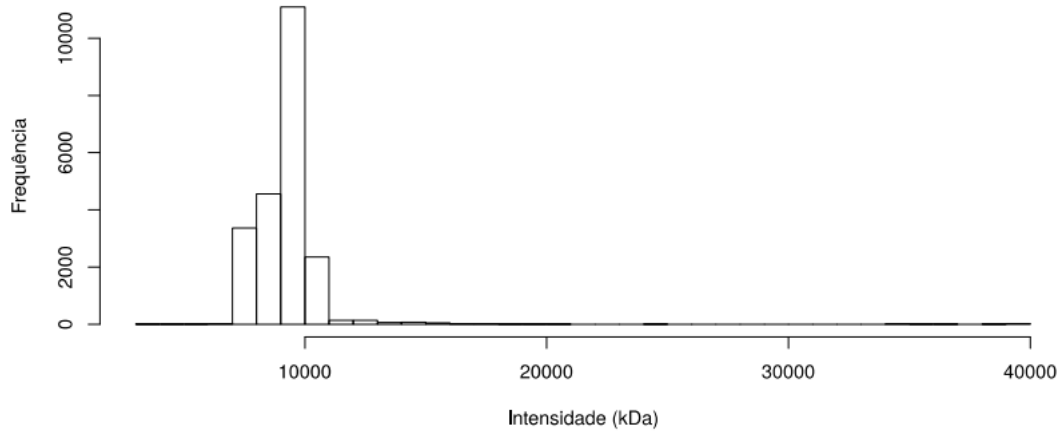
Fonte: O autor

FIGURA 51 - Histograma Proteína Ribossomal L30



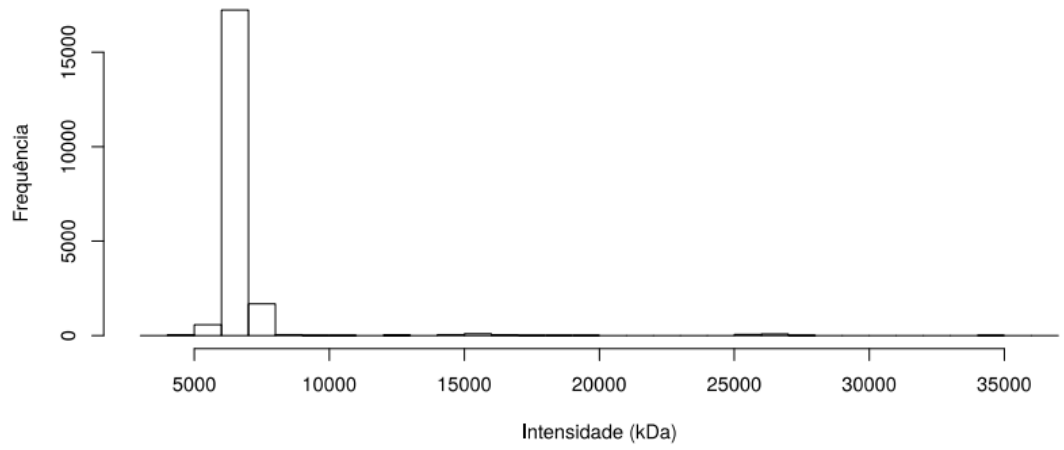
Fonte: O autor

FIGURA 52 - Histograma Proteína Ribossomal L31



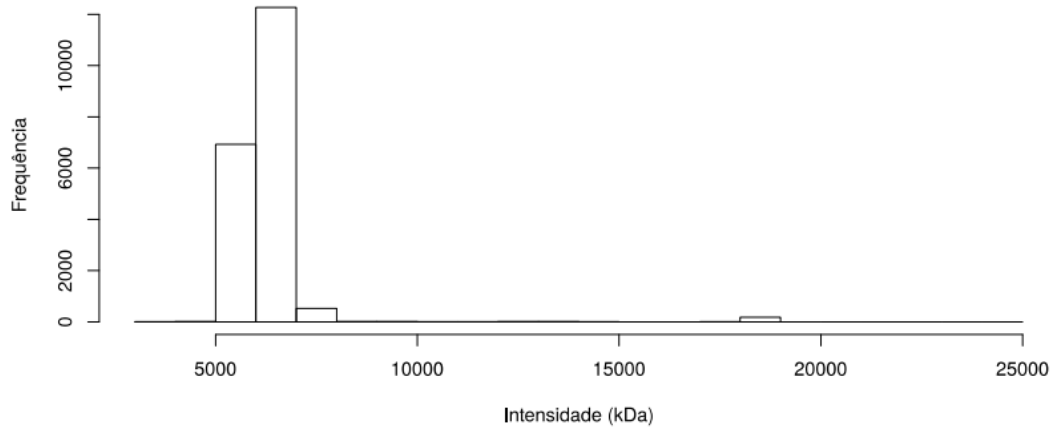
Fonte: O autor

FIGURA 53 - Histograma Proteína Ribossomal L32



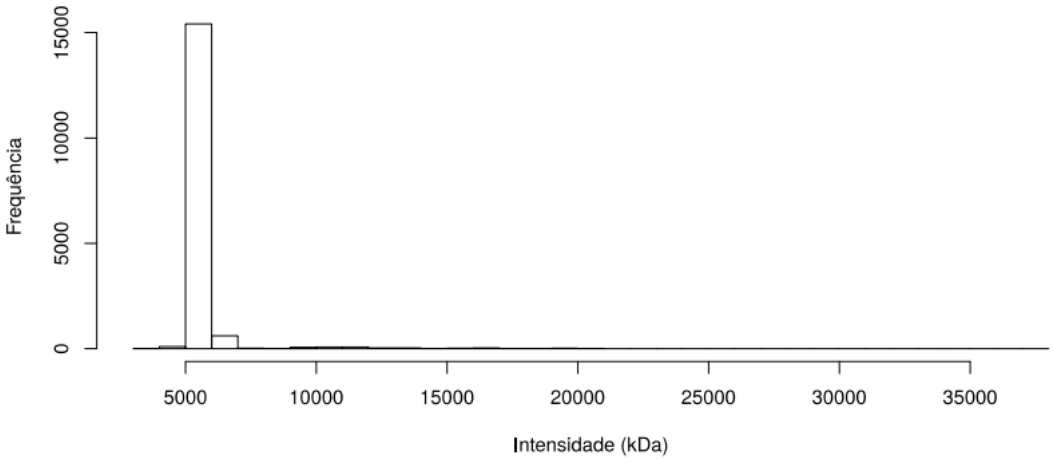
Fonte: O autor

FIGURA 54 - Histograma Proteína Ribossomal L33



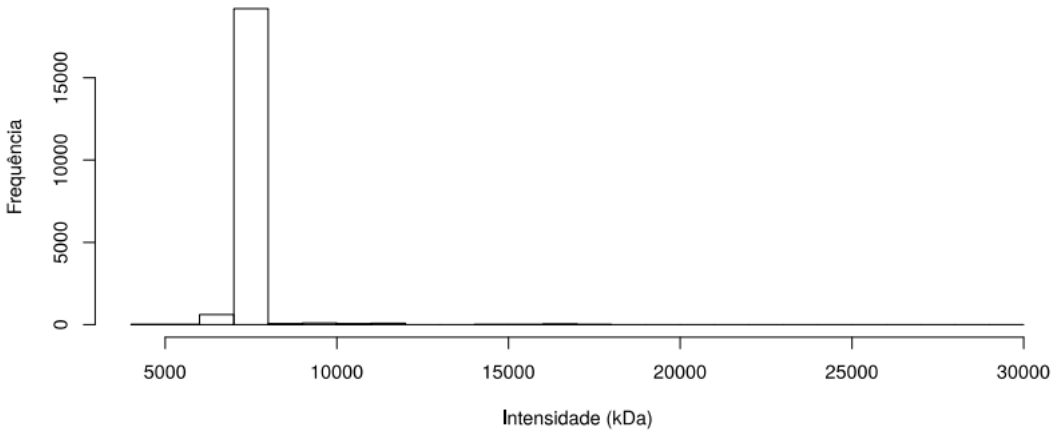
Fonte: O autor

FIGURA 55 - Histograma Proteína Ribossomal L34



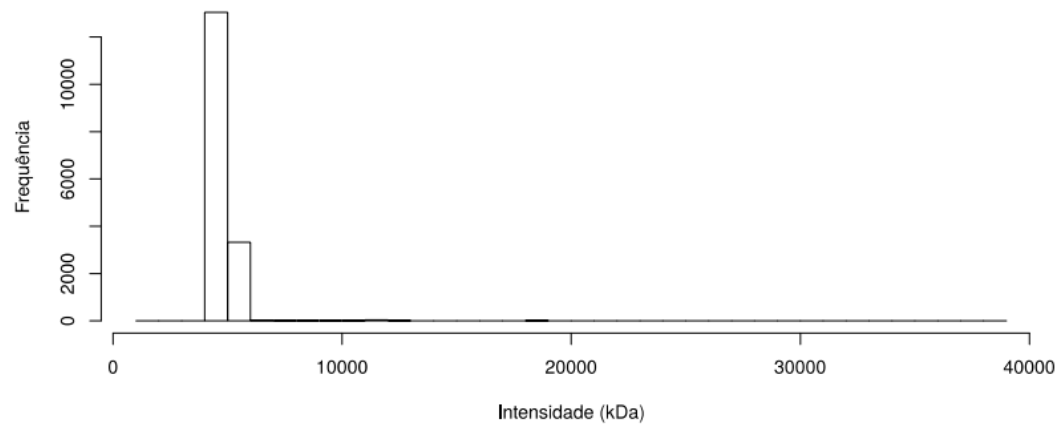
Fonte: O autor

FIGURA 56 - Histograma Proteína Ribossomal L35



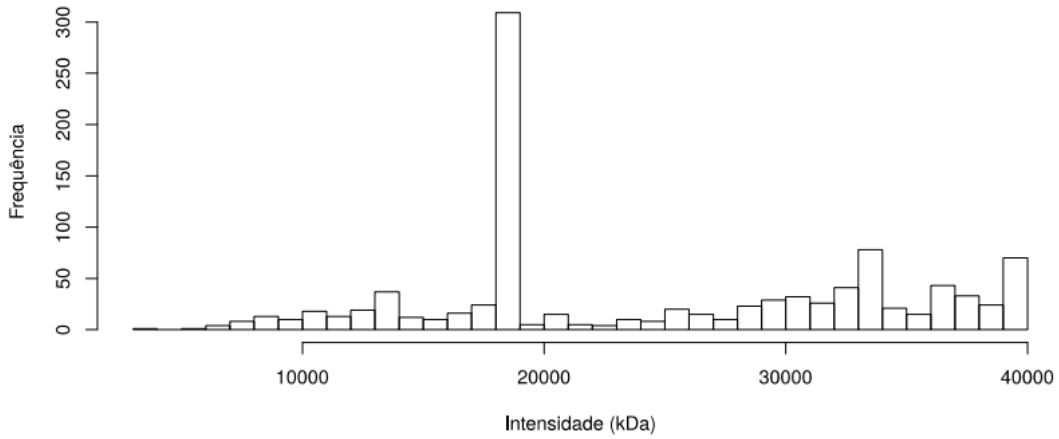
Fonte: O autor

FIGURA 57 - Histograma Proteína Ribossomal L36



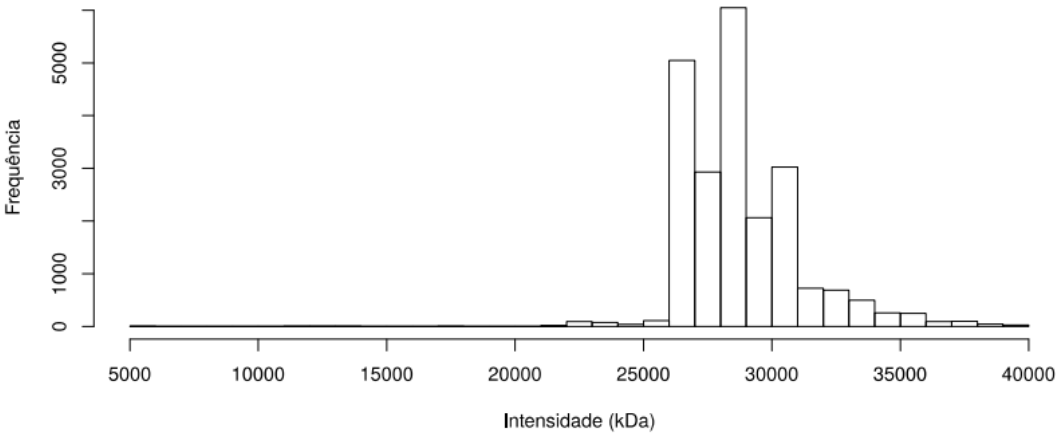
Fonte: O autor

FIGURA 58 - Histograma Proteína Ribossomal S1



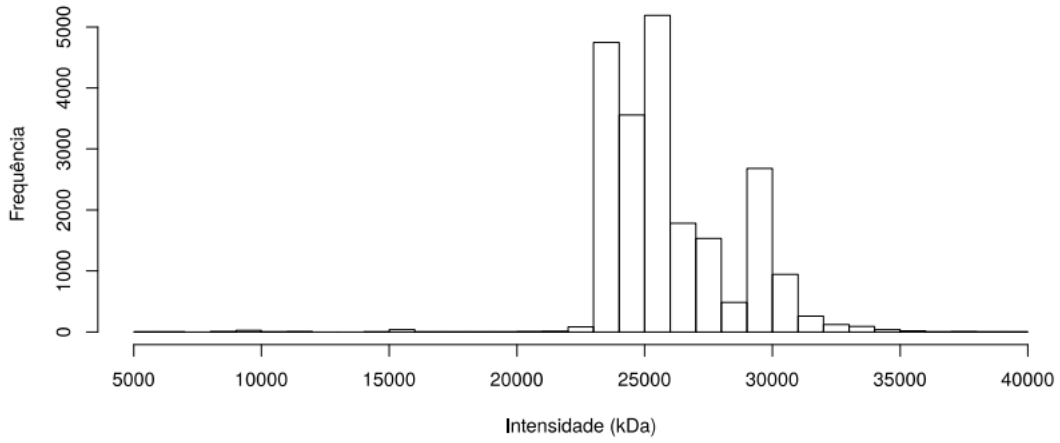
Fonte: O autor

FIGURA 59 - Histograma Proteína Ribossomal S2



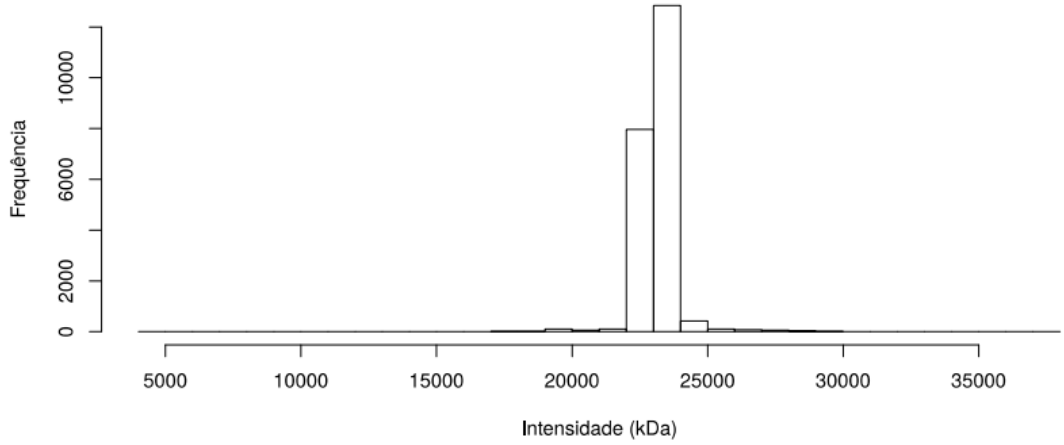
Fonte: O autor

FIGURA 60 - Histograma Proteína Ribossomal S3



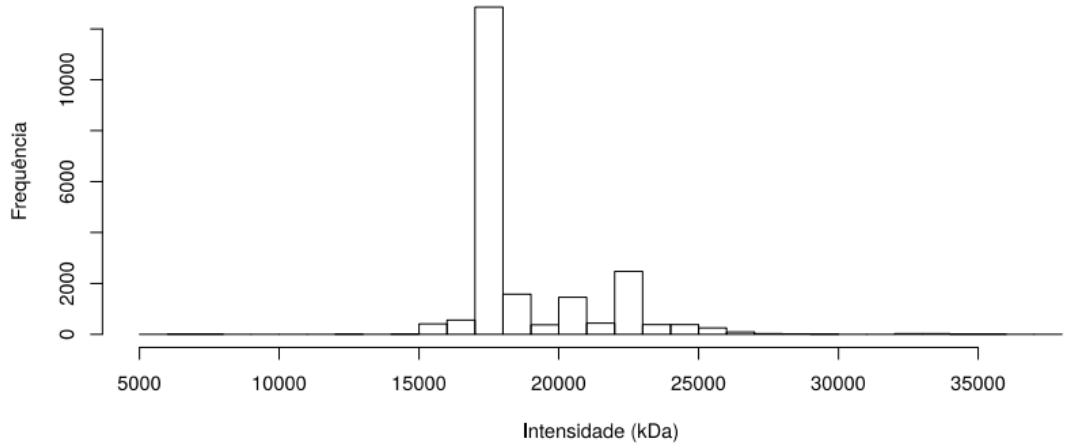
Fonte: O autor

FIGURA 61 - Histograma Proteína Ribossomal S4



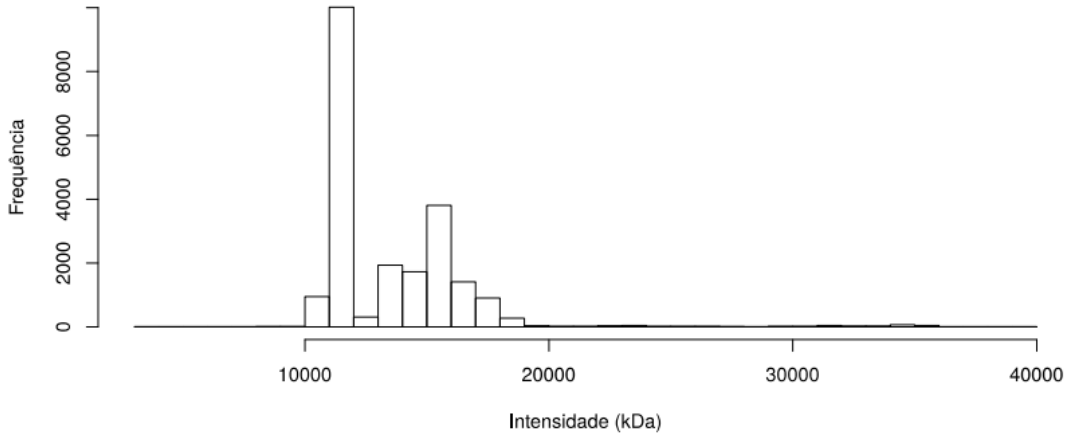
Fonte: O autor

FIGURA 62 - Histograma Proteína Ribossomal S5



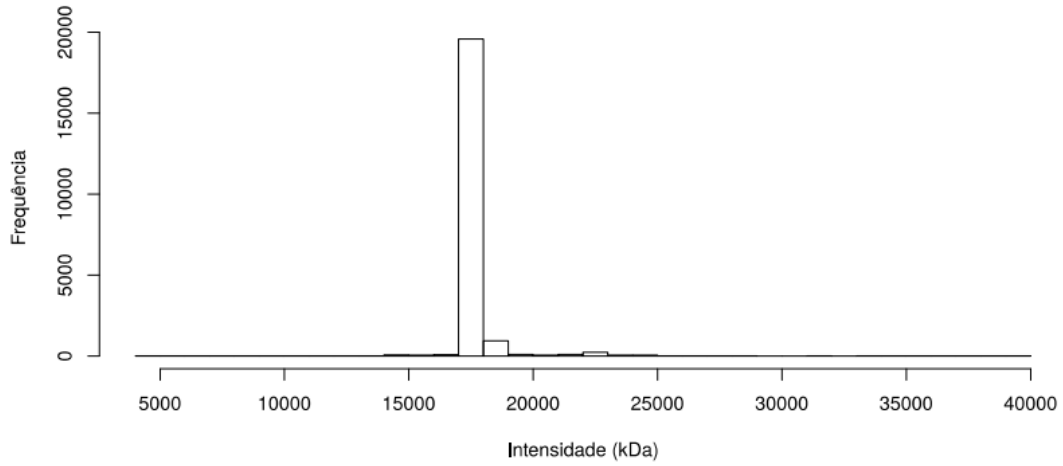
Fonte: O autor

FIGURA 63 - Histograma Proteína Ribossomal S6



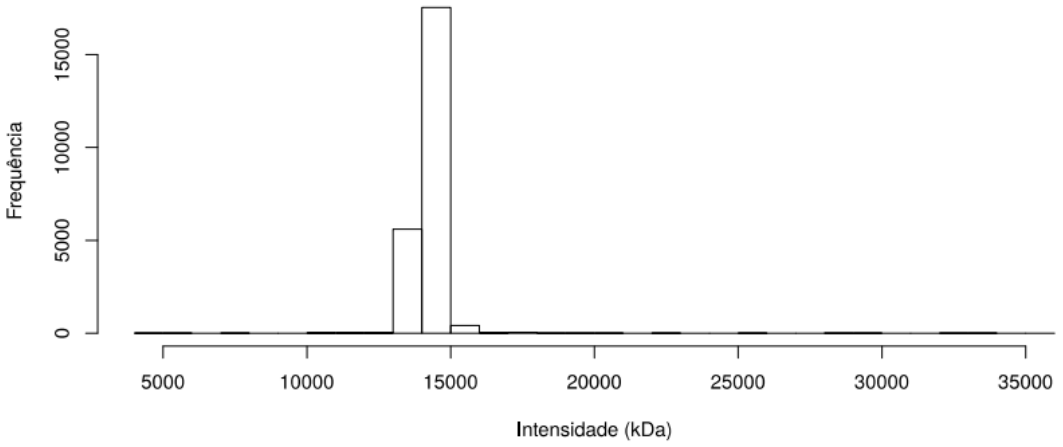
Fonte: O autor

FIGURA 64 - Histograma Proteína Ribossomal S7



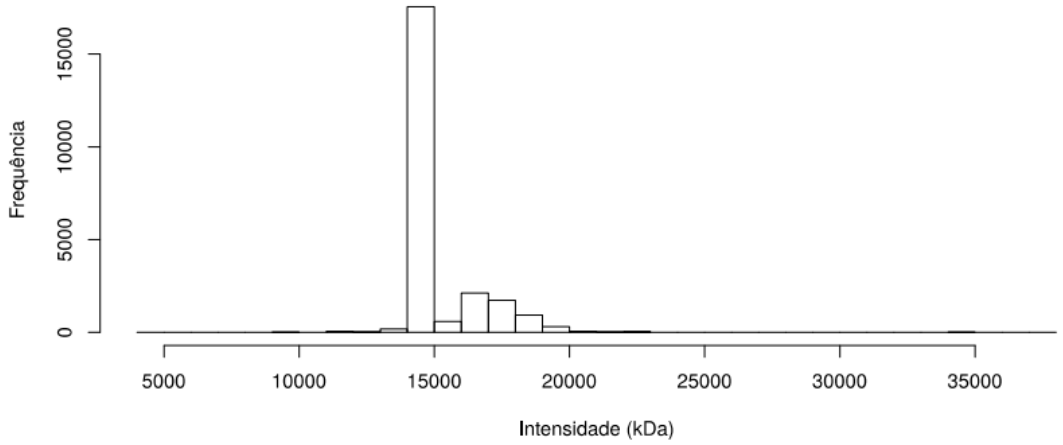
Fonte: O autor

FIGURA 65 - Histograma Proteína Ribossomal S8



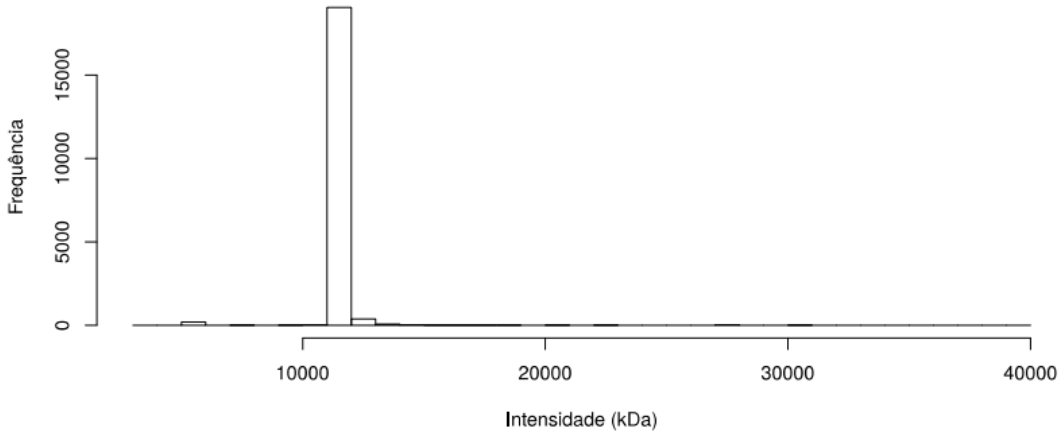
Fonte: O autor

FIGURA 66 - Histograma Proteína Ribossomal S9



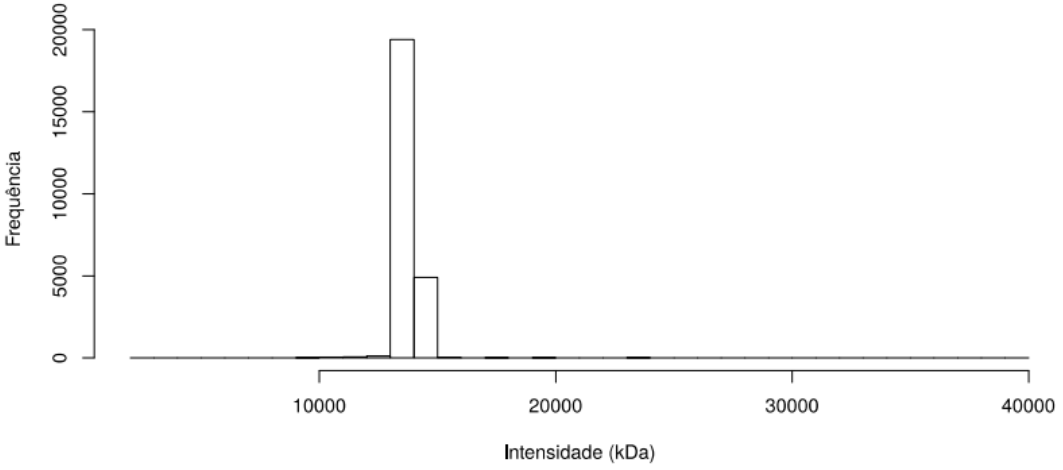
Fonte: O autor

FIGURA 67 - Histograma Proteína Ribossomal S10



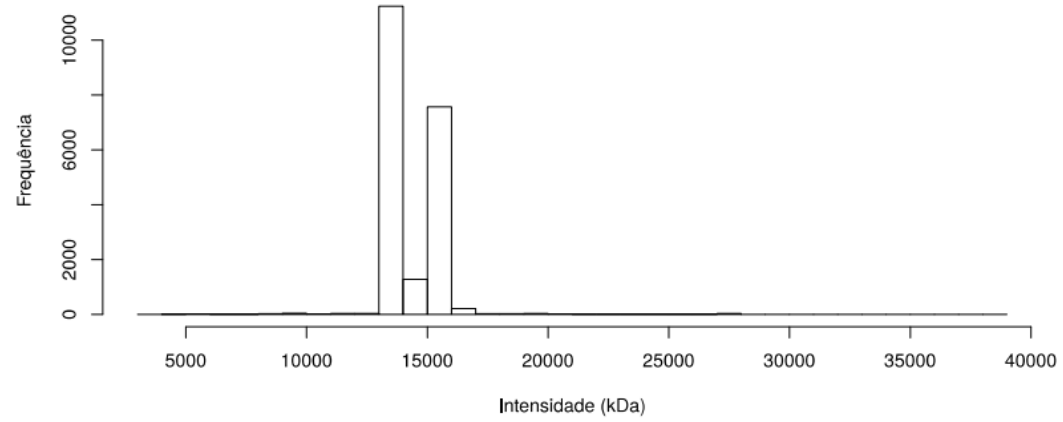
Fonte: O autor

FIGURA 68 - Histograma Proteína Ribossomal S11



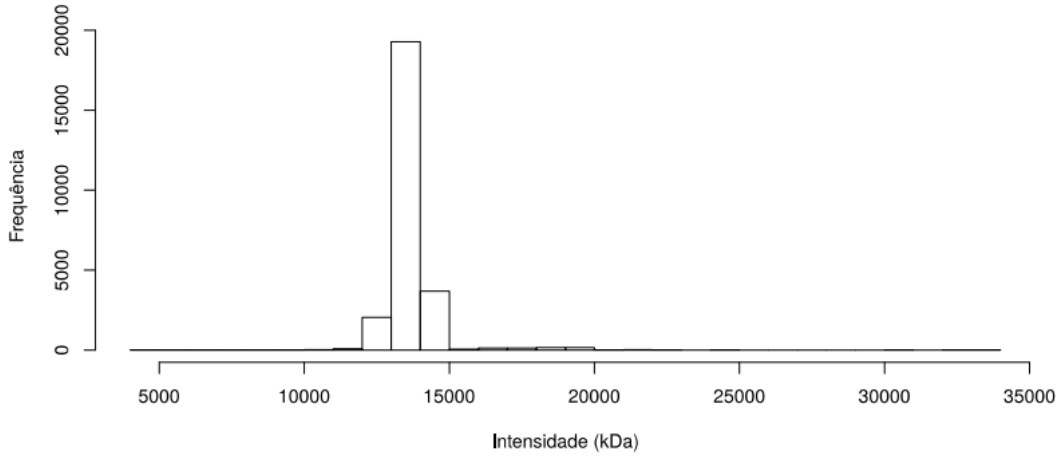
Fonte: O autor

FIGURA 69 - Histograma Proteína Ribossomal S12



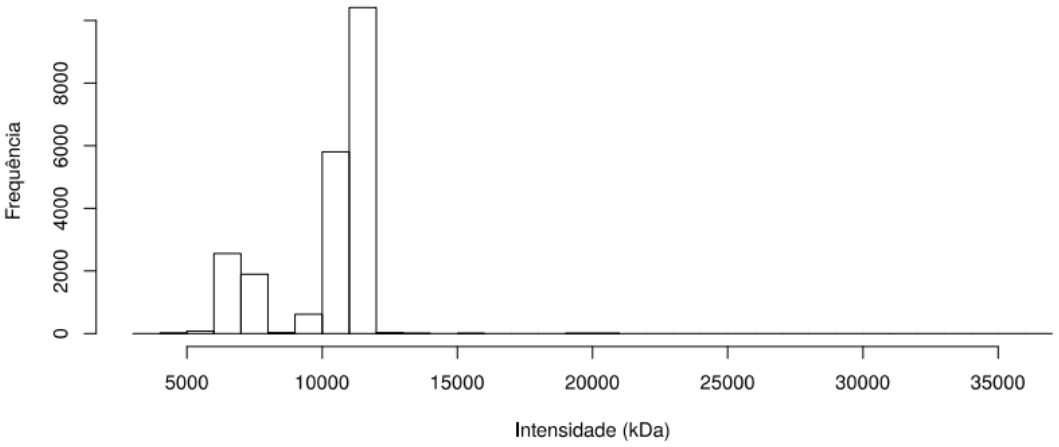
Fonte: O autor

FIGURA 70 - Histograma Proteína Ribossomal S13



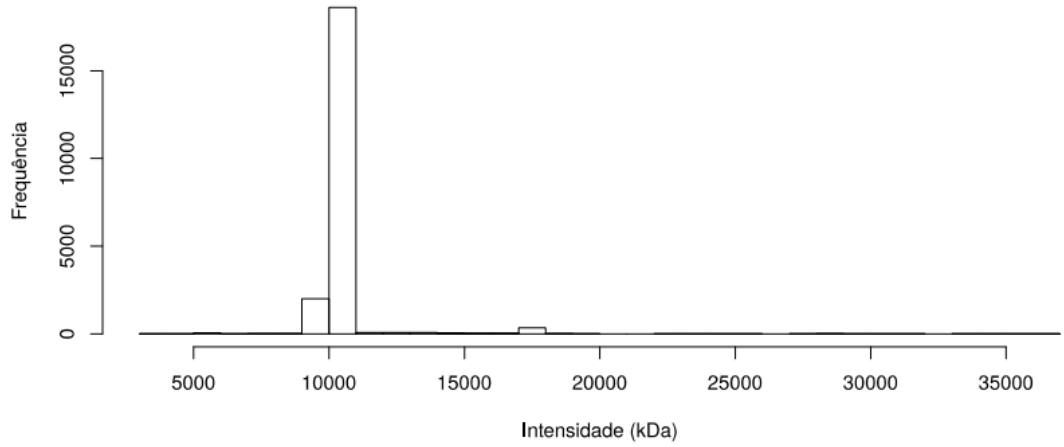
Fonte: O autor

FIGURA 71 - Histograma Proteína Ribossomal S14



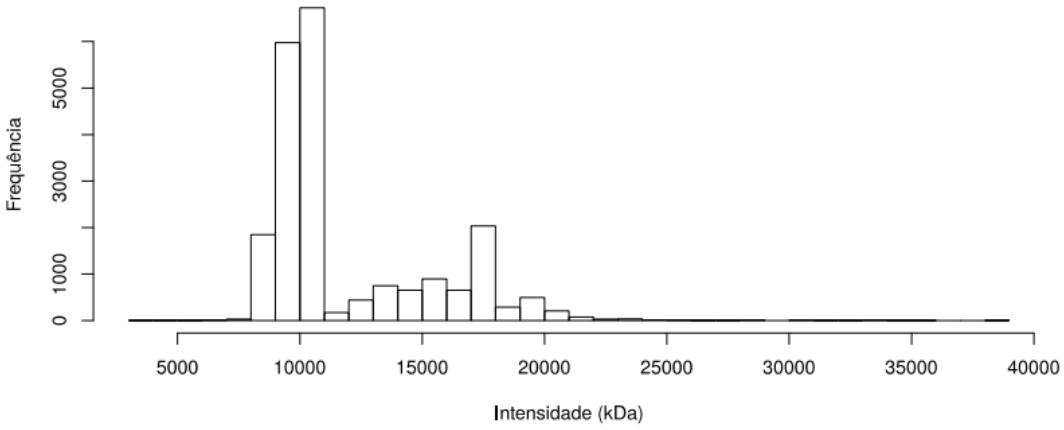
Fonte: O autor

FIGURA 72 - Histograma Proteína Ribossomal S15



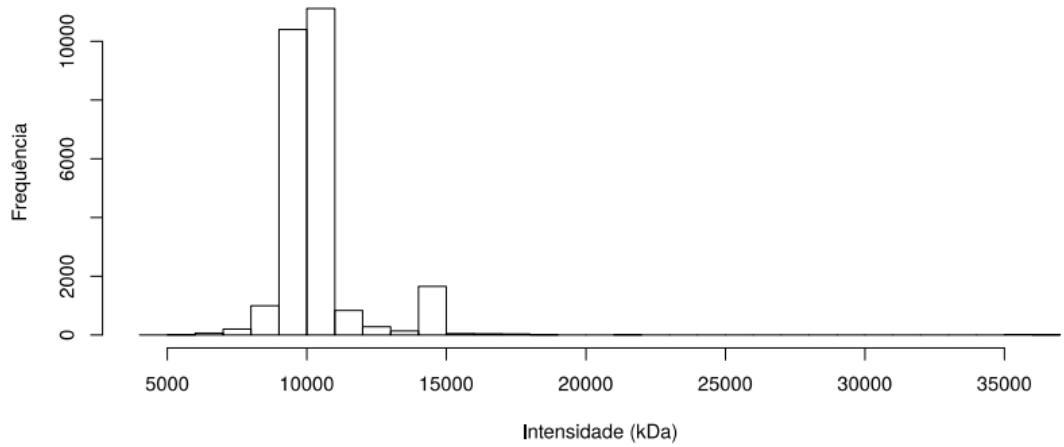
Fonte: O autor

FIGURA 73 - Histograma Proteína Ribossomal S16



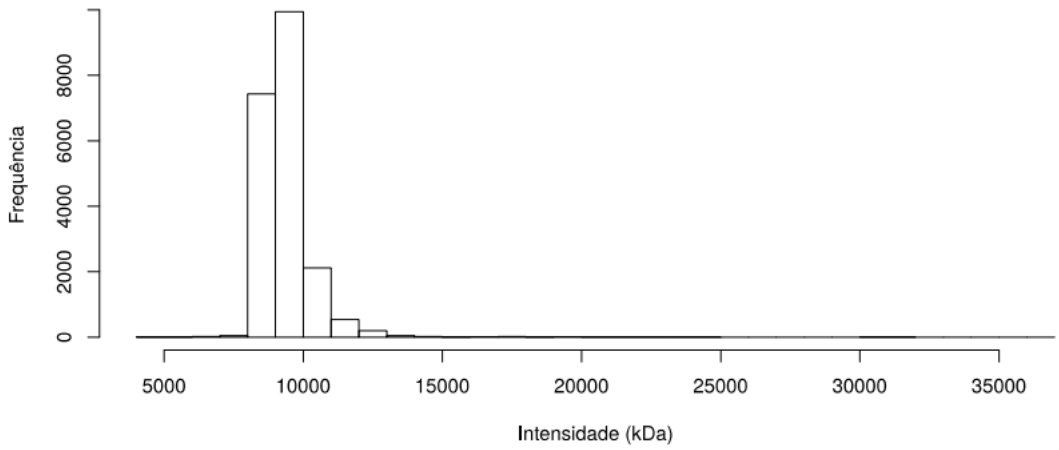
Fonte: O autor

FIGURA 74 - Histograma Proteína Ribossomal S17



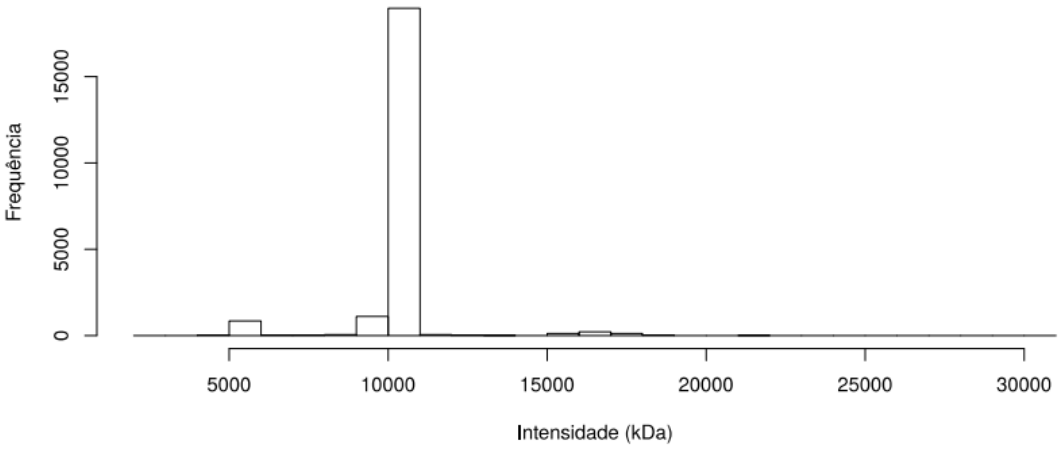
Fonte: O autor

FIGURA 75 - Histograma Proteína Ribossomal S18



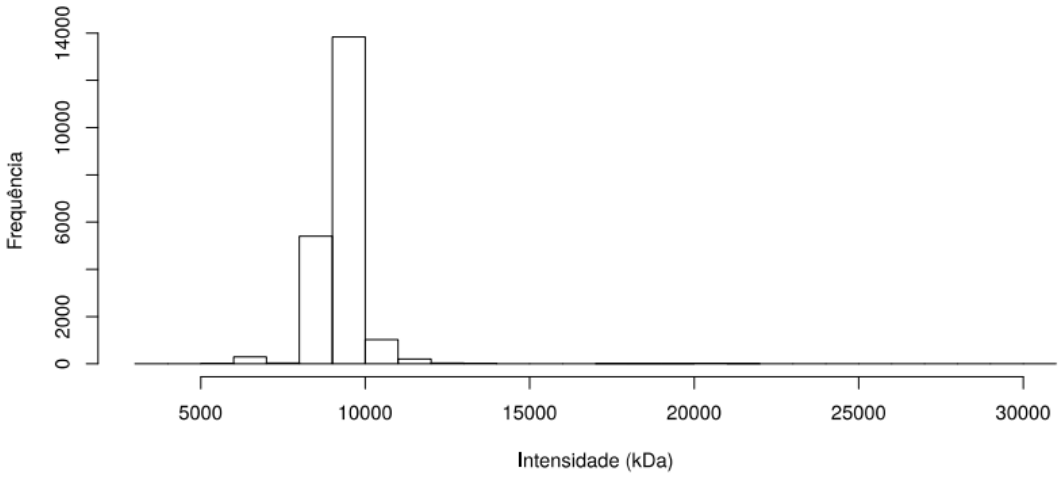
Fonte: O autor

FIGURA 76 - Histograma Proteína Ribossomal S19



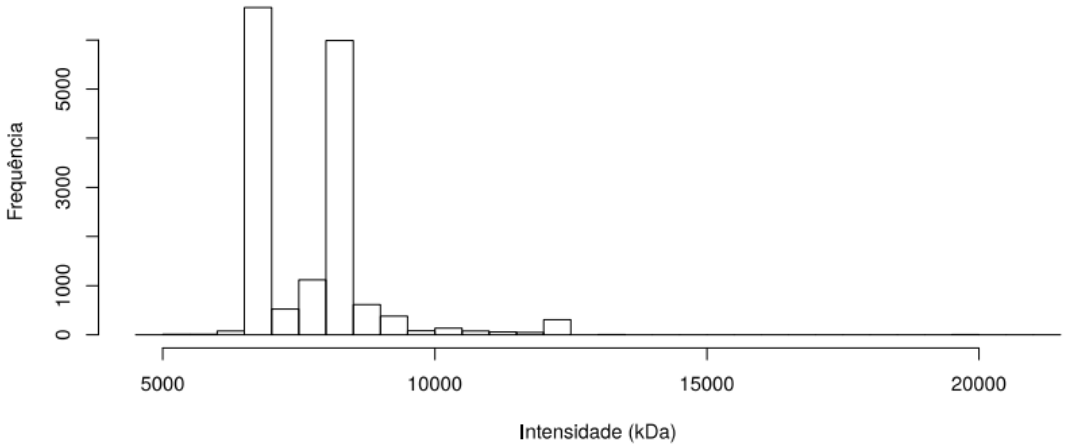
Fonte: O autor

FIGURA 77 - Histograma Proteína Ribossomal S20



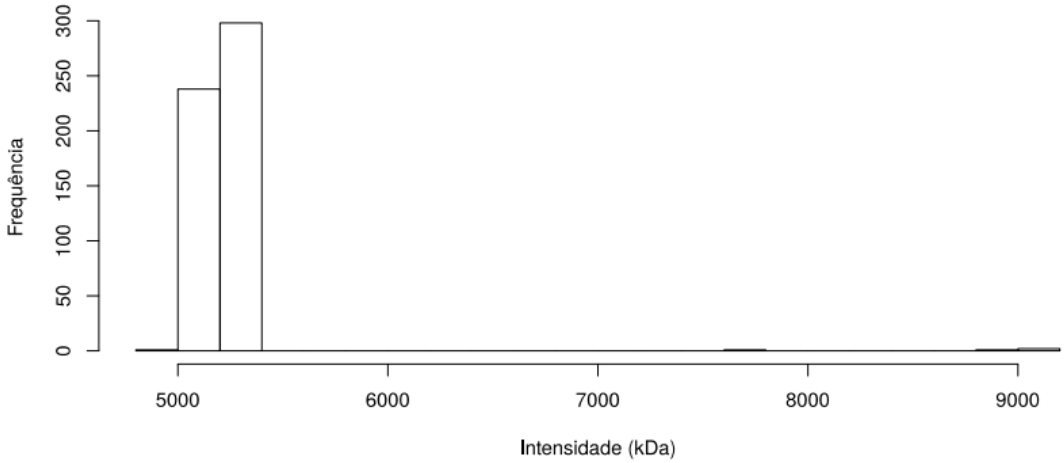
Fonte: O autor

FIGURA 78 - Histograma Proteína Ribossomal S21



Fonte: O autor

FIGURA 79 - Histograma Proteína Ribossomal S22



Fonte: O autor