UNIVERSIDADE FEDERAL DO PARÁ

INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

EDIAN FRANKLIN FRANCO DE LOS SANTOS

Abordagem computacional para a identificação de candidatos a genes housekeeping por meio de técnicas de aprendizado de máquina em dados de RNA-seq de C*orynebacterium pseudotuberculosis*

Belém do Pará
2017

Edian Franklin Franco De Los Santos

Abordagem computacional para a identificação de candidatos a genes housekeeping por meio de técnicas de aprendizado de máquina em dados de RNA-seq de *Corynebacterium pseudotuberculosis*

Dissertação de Mestrado apresentada para obtenção do grau de Mestre em Ciência da Computação. Programa de Pós-Graduação em Ciência da Computação. Instituto de Ciências Exatas e Naturais. Universidade Federal do Pará.

Orientador: Prof. Dr. Rommel Thiago Jucá Ramos

Co-orientador: Prof. Dr. Jefferson Magalhães de Morais

UFPA/ICEN/PPGCC
Belém
2017

Dados Internacionais de Catalogação - na - Publicação (CIP)
Biblioteca de Pós-Graduação do ICEN/UFPA

Franco de los Santos, Edian Franklin
Abordagem computacional para a identificação de candidatos a genes housekeeping por meio de técnicas de aprendizado de máquina em dados de RNA-seq de Corynebacterium pseudotuberculosis/ Edian Franklin de Los Santos; orientador, Rommel Thiago Jucá Ramos.-2017.
99 f.: il. 29 cm
Inclui bibliografias

Dissertação (Mestrado) – Universidade Federal do Pará, Instituto de Ciências Exatas e Naturais, Programa de Pós-Graduação em Ciência da Computação, Belém, 2017.

1. Aprendizado do computador. 2. Genes housekeeping. 3. Corynebacterium pseudotuberculosis. 4. Expressão gênica. 5. Sequenciamento nucleotídeo. I. Ramos, Rommel Thiago Jucá, orient. II. Título.

CDD – 22 ed. 006.31

Nada te turbe,
Nada te espante,
Todo se pasa,
Dios no se muda.
La paciencia
Todo lo alcanza;
Quien a Dios tiene
Nada le falta;
Sólo Dios basta...

Santa Teresa de Jesús

Para mi madre, mis sobrinos y sobrinas…

AGRADECIMENTOS

A Deus pela vida, a família e a oportunidade de poder estar no Brasil, realizando um sonho e logrando uma meta de toda a vida. Obrigado Deus por tuas abençoes.

A minha Tomasina De Los Santos, por ser meu anjo, a pessoa que sempre me apoia em todos meus sonhos e metas. Gracias mami por todo tu amor e cariño, sabes que te amo infinitamente.

A minhas irmãs Zamira, Katty, Yodernis e Águeda, por ser parte essencial em minha vida e por suas ajudas para continuar adiante cada dia. Especialmente a Águeda que me apoio desde o primeiro dia neste sonho e nunca duvido de mim.

A meus sobrinhos e sobrinhas que são alegria de minha vida.

Agradeço a minha família por sues apoio, ânimos e carinhos, meus tios, tias, primos, e primas e meu cunhado Pedro, obrigado por tanto carinho e amor.

A todas as pessoas que ao longo de minha vida têm sido essenciais em minha vida: Anouk, Risely, Miguel, Elba, Yanet, Las Hermanas Terciarias Capuchinas, Sergio, Miosotis, Lillian, Martha, Carmen Isabel, Carmen Arias, e tantos outros que deram força a minha vida.

A meu queridos companheiros e companheiras de caminho e sonho, que chegamos junto a este país com uma mala de esperanças: Melissa, Hector, Yapur e Ronaldo.

Minha querida amiga e irmã da alma e o coração Jhanier, por sua amizade, carinho e por tantos momentos vividos juntos, alegrias, tristezas e esperanças, te amo mucho mi negra.

A meus anjos brasileiros Camila e Roberto, assim como suas famílias, por seu apoio, carinho, acolhida e amor, vocês são importantes para mim. Los quiero com el alma.

A professora Iracilda Sampiao pelo apoio, carinho, acompanhamento e conselhos.

A meu orientados o Prof. Rommel Ramos, por seus acompanhamentos, conhecimentos compartilhados, conselhos e por acreditar em mim e meu trabalho. Sempre vou estar imensamente agradecido com o senhor.

Meus colegas de laboratórios, por tantas experiências vividas e conhecimentos compartilhados, obrigado desde o coração.

A Organização de Estado Americano, ao grupo COIMBRA de universidade brasileiras e minha amada Universidade Federal do Pará, por la oportunidade de poder estudar em este maravilhoso país e em este belo Estado do Pará.

Ao povo brasileiro, pela oportunidade de me formar em este pais, e pela acolhida com os braços abertos.

Desde meu coração obrigado, gracias, mil gracias...

"El Señor ha estado grande con nosotros y por eso estamos felices."

#Vivounsueño # VivoenBrasil

RESUMO

Os genes Housekeeping (HKG) ou genes de referência são necessários para a manutenção das funções celulares basais, as quais são essenciais para a sobrevivências das células. Assim, espera-se que sejam expressos em todas as células de um organismo, independentemente do tipo de tecido, estado ou condição a que está submetida a célula. Para o estudo deste tipo de genes são usadas diversificadas abordagens, uma das mais utilizadas no Sequenciadores de Nova Geração (NGS) é a *RNA Sequence* (RNA-seq), uma técnica de alto rendimento, a qual permite medir o perfil de expressão genica de um tecido ou célula isolada o organismo alvo. As análises são feitas por meio do sequenciamento do DNA complementar (cDNA) para identificar a expressão genica que estão presentes no tecido ou célula-alvo. Os HKG são usados como referências ou controle interno nas reações e experimento de *Quantitative real-time chain reaction*. Os métodos de aprendizado de máquinas são aplicados em diferentes áreas dentro da genética e genômica, permitindo a interpretação de grandes conjuntos de dados, como aqueles relacionados à expressão gênica. Uma das técnicas mais usadas são os algoritmos de agrupamento, técnica que permite definir grupos de genes com perfis de expressão similares, o que possibilita o estudo quanto à função e à interação dos genes. A *corynebacterium pseudotuberculosis,* um patógeno intracelular facultativo, foi utilizado como organismo de referência. Tal organismo infecta principalmente ovelhas, cabras, equinos, entre outros ocasionando a doença linfadenite caseosa. Para o estudo, foram utilizados os conjuntos de dados de expressão de RNA-seq das linhagens 258 e 1002 desta bactéria. Neste trabalho, é apresentada uma nova metodologia para a identificação de genes Housekeeping *in-silico* através de técnicas de aprendizado de máquina e dados de expressão genica de RNA-seq. Para a aplicação desta nova abordagem, foram utilizadas técnicas não supervisionadas de agrupamentose métricas estatísticas de avaliação e distância para o processamento e análises dos dados genômicos. Como resultado, foram encontrados 16 genes candidatos a housekeeping no patógeno pesquisado, que apresentam fortes indícios de estabilidade e expressão constante, caraterísticas de possíveis genes housekeeping.

Palavras chaves — Housekeeping, RNA-seq, Clustering, Aprendizado de máquina, Sequenciadores de Nova Geração (NGS), distancia Euclidiana, Corynebacterium pseudotuberculosis

ABSTRACT

Housekeeping genes or reference genes are required for the maintenance of basal cell functions, which are essential for maintaining a cell. Thus, they are expected to be expressed in all cells of an organism, regardless of the type of tissue, status or condition to which the cell is exposed. For the study of this type of genes are used diverse approaches, one of the most used in Next Generation Sequence is the RNA Sequence, a high-throughput technique, which allows to measure the profile of genetic expression of a target tissue or cell Isolated. The analyses are performed by sequencing the complementary DNA to find out the transcription mechanisms that are present in the target tissue or cell. Machine learning methods are applied in different areas of genetics and genomics, allowing the interpretation of large datasets, such as those related to gene expression. One of the most used techniques is the clustering algorithms, a technique that allows defining groups of genes with similar expression profiles, which allows the study of the function and interaction of genes. For the identification of housekeeping gene candidates with ML technique, *corynebacterium pseudotuberculosis*, an intracellular pathogen, was used as a model organism. This organism mainly infects sheep, goats, horses, among others causing the *Caseous lymphadenitis* disease, For the study, the datasets of RNA-seq expression of strains 258 and 1002 of this bacterium were used. In this work, presented a new approached for the identification of housekeeping genes in-*silico*, through machine learning techniques and genetic data of RNA-seq. For the application of this new approach, we used unsupervised clustering techniques, statistical and distance metrics for the evaluation, processing and analysis of the genomic data. Thus, 16 candidate genes for housekeeping were found in the pathogen, which show strong indications of stability and constant expression, indicating that housekeeping genes may be possible.

Keywords - Housekeeping, RNA-seq, Clustering, Machine Learning, Next Generation Sequence, Euclidean distance, Corynebacterium pseudotuberculosis

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE SIGLAS

A: Adenina

AM: Aprendizado de máquina

C: Citosina

cDNA: DNA complementar

CV: Coeficiente de variação

dCV: Diferença do Coeficiente de variação

DNA: Ácido Desoxirribonucleico

G: Guanina

HKG: Genes Housekeeping

KDD: Descoberta de conhecimento em dados (*Knowledge Discovery Data)*

MFC: *Máximo Change Fold*

mRNA: RNA mensageiro

NGS: Sequenciadores de Nova Geração

RNA: Ácido Ribonucleico

RNA-seq: *RNA Sequence*

RT-PCR: Reação em cadeia de polimerase em tempo real **(***Real Time Polymerase Chain Reaction*)

SVM*:* Maquina de vectores de suporte

T: Timina

U: Uracila

Weka: Waikato Ambiente para análise de conhecimento (*Waikato Environment for Knowledge Analysis*)

SUMARIO

1. INTRODUÇÃO

A genética é a área da biologia responsável pelo estudo da herança e os processos relacionados com a transmissão gênica (SNUSTAD e SIMMONS, 2010). Esta área das Ciências Biológicas poder ser considerada multidisciplinar, na qual se podem aplicar conhecimentos da Física, Química, Biologia, Computação, entre outras. A Computação foi, dentre as esferas do conhecimento acima citadas, o campo que mais interagiu com a Genética nos últimos anos, a partir do desenvolvimento de processos computacionais, algoritmos e sistemas que permitem manipular dados biológicos, como informação da anotação de genes, as sequências de DNA, além da automação de processos, dando início à Biologia Computacional, que contempla a bioinformática (POLANSKI e KIMMEL, 2007).

Os Sequenciadores Nova Geração (NGS) têm revolucionado a genética, permitindo o sequenciamento de moléculas de DNA e RNA de forma rápida e a um custo baixo. Plataformas como Illumina (Solexa) Genome Analyzer, Applied Biosystem ABI SOLiD e Ion Torrent PGM têm simplificado, facilitado e melhorado os sequenciamentos (ANSORGE, 2009). Além disso, as plataformas possibilitaram, nos últimos anos, a produção de dados genômicos e transcriptômicos, os quais aumentaram de forma exponencial, em comparação com tecnologias anteriores, sendo que a quantidade de dados está sendo duplicada a cada sete meses. Espera-se que, para o ano 2025, a área da genômica tenha uma produção de dados de aproximadamente 1 Zetabyte por ano (1.000.000.000.000.000.000.000/bytes) (STEPHENS et al., 2015).

O avanço dos métodos de sequenciamento, a partir do desenvolvimento dos Sequenciadores de Nova Geração (NGS), possibilitou sequenciar organismos com maior rapidez e eficiência (SHENDURE; JI, 2008). Uma das abordagens mais utilizadas de NGS é *RNA Sequence* (RNA-seq): uma técnica de alto rendimento, que permite medir o perfil de expressão gênica de um tecido ou organismo alvos de estudo. As análises são feitas por meio do sequenciamento do DNA complementar (cDNA), com intuito de descobrir os mecanismos de transcrição que estão presente no tecido ou organismo-alvo (KNIGHT et al., 2015). A abordagem de RNA-seq fornece várias vantagens com relação a outras tecnologias de análise de expressão: pouco ruído, menor quantidade de vies e apresenta um custo menor para a sua obtenção. As análises dos dados de expressão por RNA-seq, permite descobrir novos transcritos, promotores e isoformas (KUKURBA; MONTGOMERY, 2015; ZYPRYCH-WALCZAK et al., 2015).

Outra abordagem que permite o estudo e análise da expressão do genes dentro de uma organismos especifico é a *Quantitative real-time chain reaction* PCR (qPCR), um método de escolha para a análises de expressão de genes específicos, devido à elevada sensibilidade, especificidade e simplicidade prática (CARVALHO et al., 2014a). As análises que utilizam esta técnica requerem o uso genes housekeeping (HKG) de referência. Os genes HKG ou genes de referência são requeridos para a manutenção das funções celulares basais, as quais são essenciais para manutenção de uma célula, independentemente de sua função específica nos tecidos ou organismos. Assim, espera-se que sejam expressos em todas as células de um organismo, independentemente do tipo de tecido, estado ou condição à que esteja submetida a célula (EISENBERG; LEVANON, 2013).

Devido a estas características, os genes housekeeping são usados como controle interno em diferentes pesquisas de análises de expressão (DHEDA et al., 2004). Apesar disto, existem relatos que demostram que os níveis de expressão transcriptomas destes genes podem variar em diferentes condições (JAIN et al., 2006). Os genes housekeeping constituem uma forma de controle e de normalização da expressão gênica para experimentos em RT-qPCR (EISENBERG; LEVANON, 2013; ROCHA; SANTOS; PACHECO, 2015a)

Nos últimos anos, a identificação e análises de HKG têm chamado a atenção de muitos cientistas e isto tem levado ao desenvolvimento de várias pesquisas, que têm como objetivo a identificação deste tipo de genes para diferentes organismos por meio de análises de bancada. A maioria dos experimentos avalia a estabilidade da expressão dos genes selecionados nas diferentes condições de testes a que foram submetidos (ROCHA; SANTOS; PACHECO, 2015a). Muitas destas pesquisas são realizadas por meio de experimentos de laboratório, o que implica gasto de tempo para sua preparação e de recursos econômicos.

Um patógeno que nos últimos anos tem sido alvo de pesquisa nas áreas da procura de genes housekeeping ou de referência é o *Corynebacterium pseudotuberculosis*, uma bactéria intracelular facultativa, que infecta principalmente causador da doença linfadenite caseosa ovelhas, cabras e lifadenite ulcerativa em equinos, que causa perdida econômicas devido à diminuição da produção de carne, leite e lã. Este patógeno pode ser classificado em dois biovar: *ovis* e *equi*, baseado em sua capacidade de reduzir o nitrato. As cepas que são isoladas de ovinos e caprinos são primariamente classificadas como *ovis* e são nitrato negativo; por outro lado, os isolados de carvalos e bovinos são, em muitas situações, capazes de reduzir o nitrato e, em seguida, são classificados no biovar *equis* (CARVALHO et al., 2014a). A bactéria apresenta

uma alta prevalência em gados de países como a Nova Zelândia, Estados Unidos, Canadá e Brasil, sendo que este último tem reportado uma alta incidência da doença (RUIZ et al., 2011).

Para melhor compreensão da bactéria, vários estudos de expressão genômica e transcriptômica com *Corynebacterium pseudotuberculosis* estão sendo realizados para entender sua virulência e seu estilo de vida. O conhecimento do genoma da bactéria vai permitir entender os fatores de virulência e o descobrimento de antígenos, que podem ser usados para o desenvolvimento de vacinas e métodos de prevenção da doença. (RUIZ et al., 2011; SOARES et al., 2013a).

Os dados de expressão gênica refletem os níveis de atividade de vários genes em paralelo sob diferentes condições bioquímicas (BOLSHAKOVA; AZUAJE, 2003). O estudo e análises de dados de expressão genômicas e transcriptômica envolvem uma quantidade diversificada de abordagens, que têm como objetivo obter informações relevantes e úteis das sequências (STEPHENS et al., 2015). Uma das abordagens mais utilizadas é a da aprendizagem de máquina (AM). A AM fornece para a genética e a bioinformática ferramentas, modelos, algoritmos e estratégias que possibilitam a manipulação dos dados, para a obtenção e descobrimento de padrões, conhecimentos e informações dos dados genômicos. (POLANSKI, ANDRZEJ ; KIMMEL, 2007).

Várias pesquisas relacionadas à identificação de genes HKG foram desenvolvidas utilizando abordagens de AM com técnicas de agrupamentos e classificação. Em Dong et al. (2011), é apresentado um método para a predição de genes housekeeping, utilizando como parâmetro a transformada de Fourier em dados de séries temporais de expressão gênica, em conjunto com máquina de vectores de suporte (SVM), identificando 510 HKG em dados genômicos de humanos. Já em De Ferrari; Aitken (2006), é apresentado um método de classificação de HKG baseado em dados de caraterísticas físicas e funcionais, usando como atributos o comprimento do éxon e a medida de compacidade da cromática, com o algoritmo de classificação Naive Bayes, para a obtenção de novos candidatos a HKG em humanos. Estes trabalhos abrem um caminho de possibilidades para a pesquisa de identificação de HKG em outros organismos, como os procarióticos, a partir do uso de AM.

Um dos métodos de aprendizado de máquina mais interessantes para a exploração e extração de conhecimentos de dados de expressão genômica e transcriptômicos são os algoritmos para análise e descoberta de agrupamentos (*Clustering*). Esta técnica pode ser

definida como o processo de agrupamento de um conjunto físico ou abstrato de objetos, no qual os objetos ou classes similares estão em um mesmo grupo (cluster) e são diferenciados de outros grupos com características diferentes (HAN e KA, 2001). Nos dados de expressão gênica, a técnica de agrupamento permite definir grupos de genes com perfis de expressão similares e módulos de regulação gênicas, o que possibilita o estudo quanto à função e à interação dos genes (SI et al., 2014).

Os algoritmos de agrupamentos utilizam diferentes metodologias ou técnicas para a descoberta dos agrupamentos, que são aplicados segundo o tipo de dados e os objetivos da pesquisa. Entre os métodos de agrupamentos temos: particionados, hierárquico, baseado em densidade; redes neurais, método baseado em modelos e em grade. Cada tipo responde a uma caraterística diferente de dados ou a um objetivo de problema que está sendo pesquisado (DALTON; BALLARIN; BRUN, 2009; HAN, JIAWEI; MICHELINE, KAMBER; JIAN, 2011)

Para assegurar a qualidade dos agrupamentos que são produzidos pelo algoritmos, são utilizadas técnicas de avalição de clusters, as quais procuram o conjunto de grupos que melhor se adapta à partições naturais (RENDÓN et al., 2011). As métricas podem ser classificadas em três tipos: a) avaliação interna: que avalia conectividade, separação e robustez dos grupos - e para isto utiliza a informação dos dados b) avalição externa, baseada nos conhecimentos prévios acerca do conjunto de dados c) avaliação relativa, baseada na comparação das partições realizada por um algoritmo com diferentes parâmetros (BRUN et al., 2007). A utilização de técnicas de validação permite obter cluster com uma grande significância biológica e natural (BOLSHAKOVA; AZUAJE, 2003).

Pesquisas realizadas indicam que os genes housekeeping mostram fortes agrupamentos em diferentes tecidos humanos (LERCHER; URRUTIA; HURST, 2002), situação que pode ser similar para outros organismos, como os procarióticos. Com base nesta premissa, espera-se que a aplicação desta técnica de clustering em dados de expressão gênica permita definir agrupamentos de genes com perfis de expressão similares, o que possibilita o estudo quanto à função e à interação dos genes, assim como a identificação de novos candidatos a genes housekeeping a partir da proximidade com outros HKG validados (SI et al., 2014).

A exploração para a procura de candidatos a genes de referência ou housekeeping - por meio de técnicas de AM e dados de expressão gênica - abre a possibilidades de obter informação

relevante sobre este tipo de gene e assim identificar e criar abordagens e modelos que permitam trabalhar com esses elementos genéticos a partir de técnicas *in-silico*, o que levaria a uma redução de custos e tempo nas pesquisas que estão sendo desenvolvidas nesta área.

Este trabalho apresenta uma abordagem computacional que permite a identificação de candidatos a HKG, em dados de RNA-seq, a partir de técnicas de AM de agrupamentos (clustering) e métricas de distância. Para a implementação e testes, foram usados dados de expressão de *Corynebacterium pseudotuberculosis* das linhagens CP258 e CP1002, as quais foram submetidas a quatros diferentes tipos de estresse.

1.1 Perguntas da pesquisa

Esta pesquisa procura responder às seguintes perguntas:

- É possível identificar candidatos a genes housekeeping com poucas variações, por meio de dados de expressão gênica utilizando técnicas de agrupamentos?
- As técnicas de agrupamento podem geram agrupamentos significativos e funcionais que respondem aos perfis gênicos do microrganismo de referência?
- Os genes identificados como candidatos a housekeeping por meio desta abordagem apresentam um grau de estabilidade nas diferentes condições a que foram submetidos?

1.2 Identificação do problema

Na atualidade, métodos e técnicas para a identificação, comprovação e validação dos genes HKG implicam investimento de considerável quantidade de tempo e recursos econômicos nos processos de planejamentos, preparação e execução dos experimentos que são realizados na bancada nos laboratórios. Além disso, são necessárias técnicas específicas, como RT-qPCR, para identificar este tipo de genes (ROCHA; SANTOS; PACHECO, 2015b). Devido à importância destes genes para estudo e compressão dos organismos, é preciso o desenvolvimento e a construção de novas abordagens e métodos que permitam identificação rápida e simplificada de candidatos a HKG.

Laboratórios e grupos de pesquisas pequenos, em muitas ocasiões, não contam com equipamentos necessários para fazer experimentações para a correta identificação de candidatos. Pelo que uma abordagem *in-silico*, para a identificação de candidatos a HKG a partir de dados de expressão, o que pode promover um avanço nas pesquisas e no estudo de organismos específicos. A utilização de técnicas de aprendizado de máquina pode ajudar no desenvolvimento de modelos capazes de determinar padrões, que permitam a identificação deste tipo de genes em grandes conjuntos de dados.

1.3 Motivação e justificativa

Nos últimos anos, o interesse dos cientistas no estudo dos genes de housekeeping tem aumentado exponencialmente, devido sua vital importância para a compressão do funcionamento das bactérias e outros organismos. Em Rocha, Santos, Pacheco (2015) é apresentado um gráfico (Figura 1.1) dos últimos anos a respeito das publicações relacionadas com genes de referência ou housekeeping, estes genes são importantes nos estudos de RT-qPCR, pois são usados no controle para a normalização de dados em análises de quantificação, além de reduzirem o número de resultados inconsistentes.

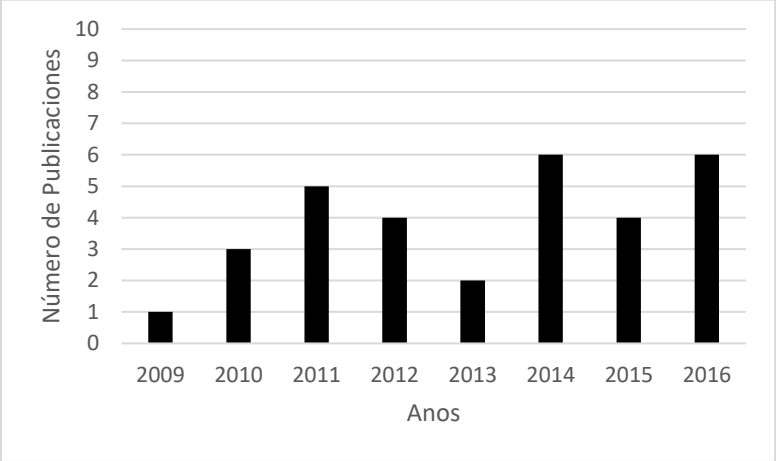


Figura 1.1: publicações relacionadas com identificação de candidatos a genes de referência ou housekeeping. (Adaptada de ROCHA; SANTOS; PACHECO, 2015b)

A RT-qPRC é a técnica mais utilizada para identificar e validar os genes housekeeping. O uso da técnicas implica um custo maior em tempo e recursos para as pesquisas que estão sendo realizadas, já que esta implica o desenho, elaboração e implementação de experimentos em bancada (ROCHA; SANTOS; PACHECO, 2015a).

A obtenção e análises de genes de housekeeping são de vital importância para a compreensão dos mecanismos e estrutura genômica dos organismos. No caso específico das bactérias, estes estudos podem ampliar o conhecimento acerca dos fatores de virulência e mecanismos de supervivência e transmissão. A compreensão desses fatores pode ajudar a reduzir o impacto dessas bactérias nas populações afetadas e auxiliar no desenvolvimento de remédios e vacinas que possam combater as doenças provocadas por elas.

O desenvolvimento de novas técnicas e abordagem que possam simplificar o processo de identificação e obtenção candidatos de genes housekeeping implicaria um grande avanço para esta área de genética, assim como tornaria mais eficiente o processo de bancada. O aprendizado de máquina fornece ferramentas e técnicas que possibilitam o desenvolvimento de uma abordagem *in-silico* por meio de dados para a identificação de candidatos a genes de referência.

Baseados no anteriormente planteado esta pesquisa tem como objetivo a construção de uma abordagem para a identificação de candidatos a genes de referência ou housekeeping a partir de técnicas de agrupamento de dados, tendo como base o uso de um conjunto de dados de expressão gênica de RNA-seq, o que podem reduzir o tempo e os custos para o desenvolvimento de estudos de identificação de possíveis HKG em diversos organismo.

1.4 Objetivos

1.4.1 Objetivo Geral

Desenvolver uma abordagem computacional para identificar possíveis candidatos a genes housekeeping em procariotos com base em dados de expressão gênica a partir de técnicas de aprendizado de máquina.

1.4.2 Objetivos Específicos

- Desenvolver uma abordagem para a identificação de candidatos a genes *housekeeping*, utilizando técnicas de agrupamento (*clustering*) e dados de RNA-seq.
- Identificar *in silico* os possíveis genes *housekeeping*, em dados de RNA-seq nas linhagens Cp258 e Cp1002 da *corynebacterium pseudotuberculosis*.
- Avaliar por meio de métricas adequadas a qualidade dos agrupamentos obtidos pela aplicação da abordagem.
- Avaliar a expressão e estabilidade dos possíveis candidatos a housekeeping identificados nos dados estudados.
- Verificar e validar por meio de uma revisão na literatura se os candidatos a genes housekeeping identificados foram descritos na literatura como possíveis genes por outras pesquisas.

1.5 Contribuições do trabalho

Dentre as principais contribuições apresentadas por este trabalho de dissertação, pode-se elencar:

- O desenvolvimento de uma nova metodologia para a identificação de candidatos a genes housekeeping por meio de uma abordagem de aprendizado de máquina;
- A apresentação um estudo de caso com dados de expressão reais de uma bactéria, com o objetivo de aplicar e verificar o funcionamento e desempenho da abordagem na identificação de genes housekeeping;
- Apresentação de uma metodologia para validação da estabilidade dos genes que são selecionados como HKG, para verificar se estes apresentam as caraterísticas mínimas para ser um gene housekeeping, através dos agrupamentos e as poucas variações de devem apresentar um genes housekeeping.

1.6 Metodologia da pesquisa

A pesquisa realizada pode ser classificada como qualitativa, baseada em dados quantitativos de expressão gênica das duas linhagens do microrganismo. Com os dados, espera-se obter uma quantidade de genes que possam ser considerados como candidatos a genes de referência ou housekeeping (SAMPIERI; FERNANDO ; BAPTISTA , 2014).

Esta pesquisa também pode ser classificada como explicativa, já que pretende explicar como podem ser identificados os genes housekeeping a partir de diferentes técnicas de agrupamentos de dados, baseadas em várias premissas que são aceitadas pela comunidade científica sobre este tipo de genes (SAMPIERI; FERNANDO ; BAPTISTA , 2014).

Os métodos e a metodologia aplicados nesta pesquisa serão detalhados em capítulo específico, no qual será detalhado o processo para obtenção dos resultados.

1.7 Estrutura do trabalho

Esta dissertação está dividida em seis capítulos, estruturados da seguinte forma:

Capítulo 2: Referencial Teórico: Exposição dos conceitos gerais que fundamentam a abordagem. Neste capítulo, podem ser encontrados os conceitos biológicos, ou seja, a

fundamentação genética e a descrição das técnicas usadas na abordagem. Também será apresentada uma visão da fundamentação computacional e de aprendizados de máquina, em que são descritos e definidos as técnicas e algoritmos implementados na abordagem. Neste capítulo, há ainda os trabalhos relacionados com a pesquisa, divididos em três grupos: trabalhos relacionados com genes housekeeping; trabalhos relacionados com *Corynebacterium pseudotuberculosis* e trabalhos relacionados com aprendizado de máquina.

Capitulo 3: Proposta e aspectos da abordagem: Neste capítulo é apresentada a descrição da abordagem, assim como as especificidades de seu funcionamento. A abordagem está dividida em três grandes processos: Pré-clustering, pré-processamento, validação dos conjuntos de dados, validação das tendências de clustering, identificação do número ótimo de agrupamento para cada conjunto de dados; Clustering, implementação dos algoritmos de agrupamentos, verificação de similaridade entre agrupamento, obtenção dos agrupamentos; Pós-Clustering: obtenção e análises das matrizes de distâncias, identificação dos possíveis candidatos a housekeeping, validação dos possíveis candidatos a housekeeping, identificação dos candidatos a housekeeping.

Capitulo 4: O estudo de caso: Neste capítulo, temos a aplicação do método no patógeno *Corynebacterium Pseudotuberculosis.* É o momento de explicar passo a passo o uso da metodologia nestes organismos, assim como as especificidades e adaptações feitas nesta implantação. Ainda temos, neste capítulo, os resultados provenientes da aplicação da abordagem e os genes obtidos como candidatos a HKG.

Capitulo 5: Discussão: Neste capítulo é feita a discussão dos resultados obtidos a partir do estudo de caso, assim como são analisadas outras questões relacionadas com a proposta.

Capitulo 6: Conclusões: Formulação das conclusões provenientes da implementação da abordagem e os resultados obtidos no caso de estudo.

2. REFERENCIAL TEÓRICO

2.1 Fundamentos

2.1.1 Ácido Desoxirribonucleico (DNA) e Ácido Ribonucleico (RNA)

O ácido desoxirribonucleico (DNA) é uma macromolécula encarregada da transmissão da informação genética de uma geração para outra. O DNA contém as informações necessárias para o desenvolvimento dos seres vivos e de alguns vírus. A função fundamental desta molécula é carregar os genes que levam as instruções necessárias para construção de novos componentes biológicos, como as proteínas e células (PIERCE, 2009; SNUSTAND; SIMMONS, 2013).

Quimicamente, o DNA é composto por subunidades repetidas chamadas de nucleotídeos (Figura 2.1). Cada nucleotídeo é composto de: um grupo fosfato, uma pentose e uma base nitrogenada. As quatro bases que compõem o DNA são adenina (A), guanina (G), timina (T) e citosina (C). As combinações destes quatro nucleotídeos codificam as informações genética de cada organismo vivo (SNUSTAND; SIMMONS, 2013).

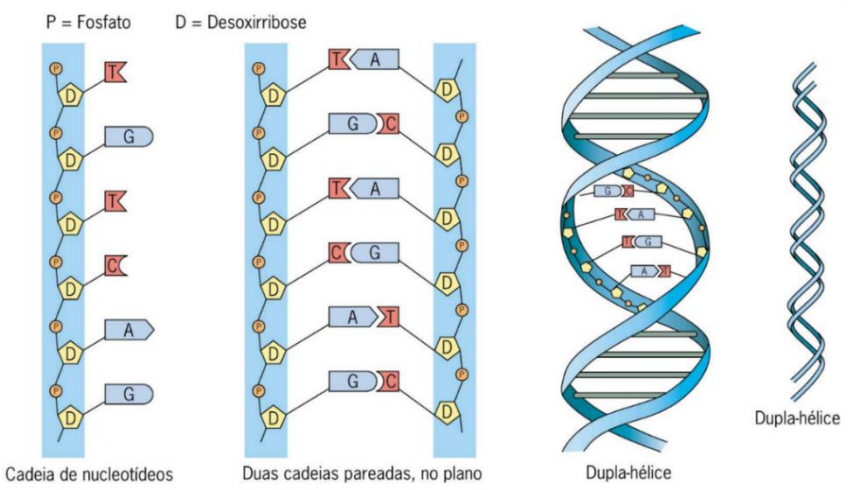


Figura 2.1: Exemplo da estrutura do DNA. (DA SILVA JUNIOR; SASSON, 2002)

Para que a informação contida no DNA exerça função biológica, é necessário que ocorram os processos de transcrição e tradução. No processo de transcrição a informação no DNA é em RNA (ácido ribonucleico) para que então, no processo de tradução, este último seja convertido em aminoácidos e forme proteínas. Para este trabalho, daremos enfoque aos ácidos nucleicos, tais quais o RNA.

O RNA é uma macromolécula presente nos organismos procarióticos e eucarióticos. Alguns vírus apresentam o RNA como portador da informação genética para o desenvolvimento deste organismo (PIERCE, 2009). A função fundamental do RNA é intermediar a transmissão da informação genética do DNA para formar as estruturas das proteínas no processo de expressão gênica. Quimicamente o RNA é composto por três bases como o DNA, está são: adenina (A), guanina (G) e citosina (C), mais uma base diferente: uracila (U) em lugar de tiamina (T) (SNUSTAND; SIMMONS, 2013). Existem diferentes tipos de RNA, como RNA ribossomal, RNA transportador e RNA mensageiro, os quais tem diferentes funções dentro dos organismos(PIERCE, 2005).

2.1.2 Genes Housekeeping

Os genes housekeeping podem ser definidos como genes que são necessários para a manutenção de funções celulares básicas que são essenciais para a existência de uma célula, independentemente do seu papel específico no tecido ou organismo. Assim, eles são expressos em todas as células de um organismo sob condições normais e situações de estresse, independentemente do tipo de tecido, do estado de desenvolvimento, ciclo celular ou um sinal externo. A caracterização do conjunto mínimo de genes necessários para sustentar a vida, tais quais os genes housekeeping, é de especial interesse para a comunidade cientifica (EISENBERG; LEVANON, 2013). Algumas pesquisas sugerem que este tipo de gene pode ter funções importantes nos processos de transcrição, tradução e sinalização celular (REUE; GLUECK, 2001).

Os genes housekeeping são usados em muitas pesquisas como controles internos em experimentos com microrganismo, como na análise de expressão gênica, por serem expressos em todas as células, já que são necessários para a sobrevivência destas. Um ponto importante é que algumas pesquisas têm indicado que este tipo gene também pode ter variações, que dependem do metabolismo e o tipo de célula. Técnicas que permitem a quantificação do RNA mensageiro (mRNA) como RNase protetion, Norther blot, RT-PCR usam os genes como controles interno da expressão dos dados (THELLIN et al., 1999).

2.1.3 Genoma Central

É o conjunto de genes compartilhado por toda a estipes de um conjunto de genomas bacterianos (Figura 2.2). Estes estão contido no pan-genoma que é o repertório genético global de uma espécie bacteriana. Em geral o core genoma inclui todos os genes responsáveis pelos

aspectos básicos da biologia de uma espécie e suas principais características fenotípicas(MEDINI et al., 2005).

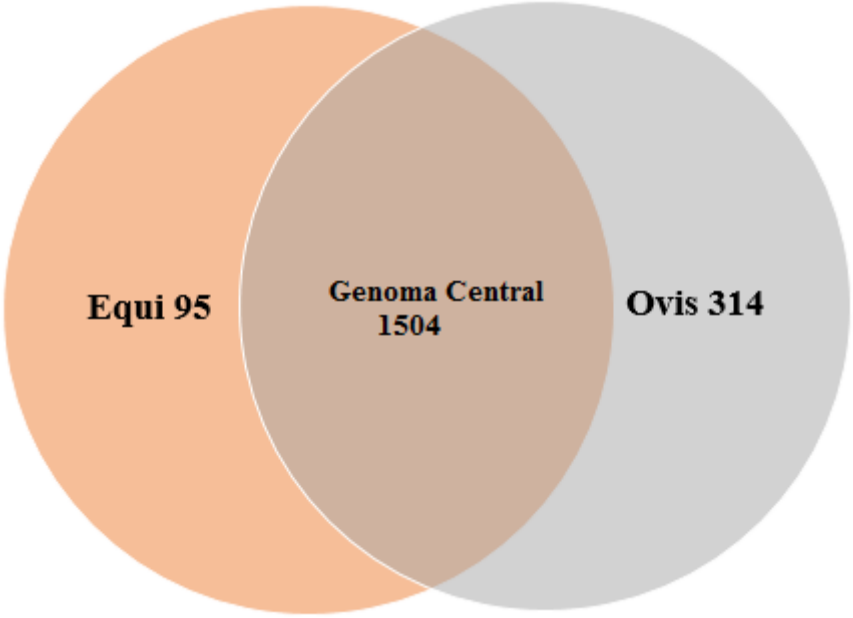


Figura 2.2: Exemplo do Core Genoma de C. Pseudotuberculosis.(SOARES et al., 2013b)-Adaptada

2.1.4 *Corynebacterium Pseudotuberculosis*

É uma bactéria grande- positiva, pleomórfica e anaeróbia facultativa da ordem dos *Actinomycetales*. É um microrganismo intracelular facultativo que pode proliferar dentro dos macrófagos (SOARES et al., 2013a). Esta bactéria é o organismo causador de doenças crônicas como a linfadenite gaseosa. Esta doença tem ocorrência mundial e acomete caprinos, ovinos, equídeos, bovinos, suínos e cervos, provocando perdas econômicas, em função da redução na produção de lã, carne, leite e do desperdício da carcaça. Há, ainda, raros relatos da ocorrência da doença em seres humanos. Nos animais a infeção pode causar abscessos externos (linfonodos) (Figura 2.3), assim como, pode comprometer órgãos internos ou infecção dos membros na forma de linfagite ulcerativa(DORELLA et al., 2006; PEPIN; BOISRAME; MARLY, 1989; SELIM et al., 2016; SOARES et al., 2013b)

As doenças causadas por *C. pseudotuberculosis* tem incidência mundial e apresentam alta prevalência em países produtores de carne como Austrália, Nova Zelândia, África do Sul, Estados Unidos, Canadá e Brasil (SOARES et al., 2013a).

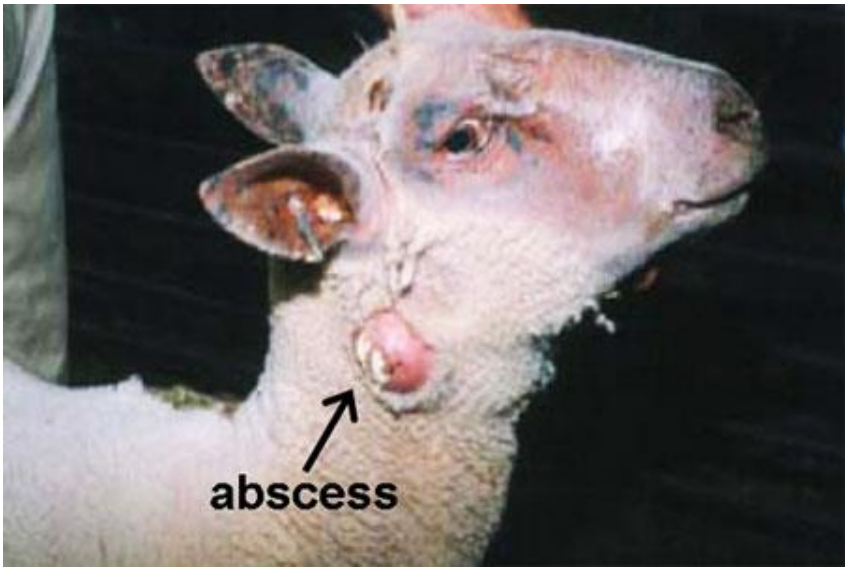


Figura 2.3: Animal infectado com *corynebacterium pseudotuberculosis*, pode-se identificar o nódulo caraterístico da doença de linfadenite caseosa, infecção da bactéria. Fonte: http://en.engormix.com/

2.1.5 Técnicas Sequenciamento de Nova Geração (NGS)

O descobrimento da estrutura e a formação das moléculas de DNA, por parte de Watson e Crick, estimulou o desenvolvimento de técnicas de sequenciamento de DNA (SNUSTAD e SIMMONS, 2010). O sequenciamento pode ser definido como o conjunto de métodos e técnicas que permitem a determinação da ordem dos nucleotídeos em uma macromolécula de DNA (HINDLEY, 1983).

As tecnologias de sequenciamento de nova geração (NGS) são a evolução de métodos de sequenciamentos mais antigos como o método de Sanger. Desde sua introdução no mercado a partir de 2005 os NGS têm tido um grande impacto nas pesquisas genômicas. Estas tecnologias são usadas para sequenciamento padrão ou para sequenciamentos e re-sequenciamentos de organismos, assim como para novas aplicações anteriormente não exploradas por sequenciadores como que utilizavam o método de Sanger. As tecnologias NGS oferecem um aumento na taxa de transferências de sequencias custo-benefício, embora as leituras (reads) seja de menor cumprimento comparada com os resultados do método de Sanger (MOROZOVA; MARRA, 2008).

Existem diferentes plataformas de sequenciamento, que usam diferentes técnicas bioquímicas para a obtenção da sequência de DNA ou DNA complementar (cDNA). Entre estas tecnologias podemos citar as seguintes:

- **454 GenomeSequencer FLX instrument**: esta plataforma está baseada no princípio de

detecção de pirofosfato, que foi desenvolvida pela companhia 454 Life Science (SHENDURE; JI, 2008)

- **Illumina (Solexa) Genome Analyzer**: esta plataforma de sequenciamento foi comercializada no ano 2006 e realiza o sequenciamento por síntese química, com nucleotídeos terminadores reversíveis para as quatro bases cada um marcado com um corante fluorescente diferentes. (ANSORGE, 2009).
- **Applied Biosystem ABI SOLiD System:** é uma plataforma de sequenciamento baseada em ligase, que foi comercializada a partir de 2007 (ANSORGE, 2009).
- **Helicos singler-molecule sequencing device, HeliScope:** Esta plataforma requer a emulsão PCR (reação em cadeia da polimerase), para a amplificação dos fragmentos de DNA(ANSORGE, 2009).
- **Ion Torrent:** aproveita o poder da tecnologia de semicondutores, detecta os prótons liberados como nucleotídeo durante o processo de polimerização de DNA. Os fragmentos de DNA com as sequencias especificas estão ligados, e são submetidos a um processo de amplificação por clonagem através da emulsão de PCR sobre uma superfície de 3micro conhecida como "Ion Sphere Particles", esta permite a detecção dos prótons por sinal que são proporcionais ao número de bases presente na amostra.(QUAIL et al., 2012)

2.1.6 Transcriptoma

Wang (2009) define este como o conjunto completo de transcritos em uma célula e suas quantidade para o desenvolvimento de uma fase específica ou uma condição fisiológica. Também pode ser definido como o conjunto completo de transcritos nas diferentes condições da célula ou tecido. Portanto, ele é o reflexo direito da expressão dos genes. O transcriptoma é essencial para a compressão do funcionamento do genoma, a formação de tecidos e os estudos dos processos de infecção produzidas por vírus e bactérias (PIERCE, 2009).

Para o estudo de transcriptoma se utilizam diferentes técnicas para a detecção, analises e tratamentos dos diferentes tipos de RNA. Atualmente, as principais técnicas utilizadas no mercado são: qPCR, Microarray e RNA-Seq (MOROZOVA, HIRST e MARRA, 2009).

A técnica Real Time PCR ou PCR quantitativa, é uma técnica precisa, eficaz e rápida para a detecção de expressão gênica. Esta técnica está baseada na técnica tradicional *Polymerase Chain Reactor* (PCR) (TAYLOR E FRANCIS GROUP, 2007).

Microarray, é uma técnica baseada na hibridação de cadeias de alvo por sondas. Ao medir a luz dos corantes florescentes após a hibridização das sondas obtém-se a abundância relativa de cada transcrito nas condições de estresse avaliadas. As plataformas capazes de processar microarray são Affymetrix e Agilent (KOGENARU et al., 2012).

2.1.7 *RNA Sequencing* (RNA-seq)

Com os sequenciadores NGS surgiu a técnica RNA-seq para a quantificação e análise de transcriptomas. A técnica funciona da seguinte forma: em geral uma amostra de RNA (total ou fraccionado) é convertida em uma biblioteca de cDNA com adaptadores ligados em ambas as extremidades. Cada molécula com ou sem amplificação é sequenciada para obter as sequencias curtas de uma extremidade (*single-end*) ou ambas (*paired-end*) (WANG; GERSTEIN; SNYDER, 2009). As sequências curtas resultantes são usadas para mapear o genoma de referência e quantificar os níveis de expressão relativos ou absolutos dos genes em estudo em diversas condições, na Figura 2.4 demostra-se o processo de montagem de um experimento baseado nesta técnica. Ela pode ser implementada em diferentes plataformas como o Illumina´s Genome Analyzer, Rocher 454 Life Science e Applied Biosystems SOLiD (KOGENARU et al., 2012)

Esta técnica de análise de transcritos tem grandes vantagens em comparação com a abordagem e técnicas anteriores, pois não precisa dos conhecimentos prévio da sequência do gene para avaliar sua expressão, tal como ocorre em técnicas que utilizam hibridização de sondas e ainda possui um custo baixo. RNA-seq é o primeiro método que permite a avaliação do conjunto total de transcritos de um organismo sob uma determina condição de estresse (WANG; GERSTEIN; SNYDER, 2009).

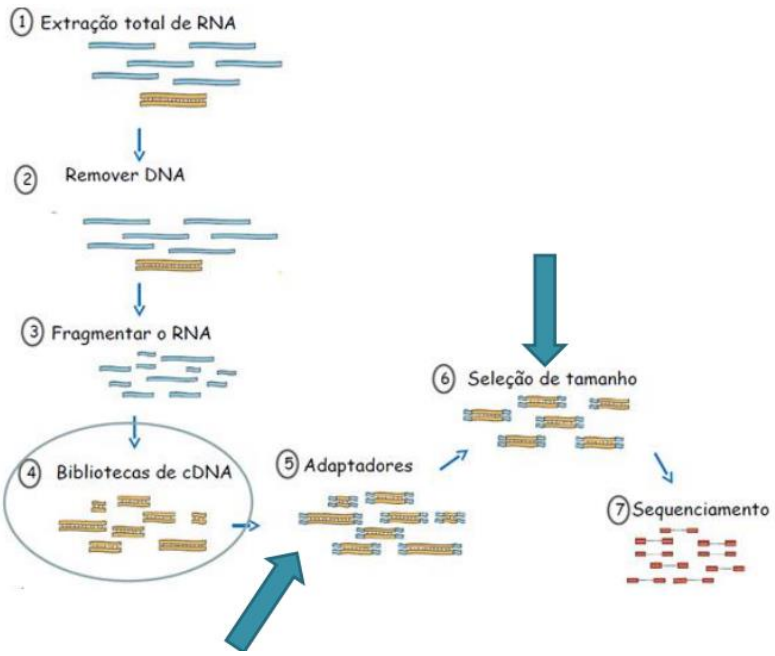


Figura 2.4: Fluxograma de um experimento de RNA-seq. (Adaptada de MARTIN; WANG, 2011)

2.1.8 Kilobase per Million Reads (RPKM)

Uma dificuldade que se enfrenta quando se trabalha com dados de sequenciamento é a grande diferença no número de leitura produzidas na execução para a obtenção das sequencias, assim como as viés técnicos que são introduzidos pelos protocolos de preparação das bibliotecas, plataformas utilizadas e composições nucleotídicas. Os métodos de normalização fornecem uma solução para este problema e permitem a comparação precisa entre os grupos de amostras (RAPAPORT et al., 2013).

Uma das medidas de normalização mais usada para os dados de RNA-seq é RPKM. O qual é calculado a partir do número de leituras mapeadas em uma região particular do gene, tamanho da região, e o número total de leituras mapeadas na sequência (WAGNER; KIN; LYNCH, 2012). Este método elimina os efeitos do comprimento e do tamanho das bibliotecas de sequenciamento(CONESA et al., 2016). Nesta normalização a representação de cada gene é dependente dos níveis de expressão de todos os outros genes (WAGNER; KIN; LYNCH, 2012).

2.1.9 Aprendizado de Máquina

Witten (2011) define a aprendizado de máquina (AM) como: “o processo de descoberta de padrões de dados. O processo deve ser automático ou semiautomático. Os padrões descobertos devem ser significativos, na medida em que trazem algumas vantagens, geralmente

econômicas". Outra definição de aprendizado é a prática de pesquisa automática de grandes bancos de dados para descobrir padrões e tendências que vão mais longe de uma simples análise. Este processo usa diferentes algoritmos matemáticos e estatísticos para segmentar os dados e avaliar as probabilidades de futuros eventos. A AM pode responder a perguntas que não podem ser abordadas a partir de uma simples consulta aos bancos de dados (ORACLE, 2008).

No seu artigo sobre o tema, Fayyad (1996) define a aprendizagem de máquina como um passo dentro do processo de KDD, cuja função é a aplicação de algoritmos de análise e descoberta de dados com o objetivo de produzir uma enumeração particular dos padrões de dados.

Os objetivos mais importantes do aprendizado de máquina são a predição e a descrição a partir dos dados recebidos como entradas. A descrição delineia as características gerais das informações no banco. A predição, realiza inferências baseadas nos dados para realizar prognósticos. Para obter estes objetivos, diferentes métodos de AM primários são utilizados, tais como:

- **Classificação**: aprender uma função que mapeia (classifica) um item ou objeto de dados em uma das várias classes predefinidas. O algoritmo procura classificar os elementos dentro das classes mais apropriadas, segundo seus atributos ou características (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).
- **Agrupamentos:** identificação de um conjunto finito de categorias ou grupos para descrever os dados. O objetivo do agrupamento é a formação de categorias a partir de um conjunto físico ou abstrato de objetos, no qual os objetos ou classes similares estão em um mesmo grupo (cluster), diferenciado de outros grupos com características diferentes (HAN e KA, 2001).
- **Associação:** é um processo que procura encontrar as relações entre as diferentes características que formam parte do conjunto (HAN e KA, 2001).
- **Regressão:** aprender uma função com o intuito de mapear um conjunto de dados, para que haja a predição de variáveis de valores reais e o descobrimento das relações funcionais entre as variáveis (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).
- **Modelo de dependências**: procura um modelo que descreve dependências significativas entre variáveis (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).
- **Detecção de mudanças e desvios:** descobre as mudanças mais significativas

nos dados a partir de valores medidos e normativos (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

A aplicação dos métodos anteriores por meio de algoritmos específicos em AM requer um processo formado por três componentes, que vão formar qualquer algoritmo a ser utilizado: 1) representação de modelos; 2) avaliação de modelo e 3) modelo de pesquisas.

- **Representação de modelo:** é a linguagem usada para descrever os padrões descobertos pelo algoritmo.
- **A avaliação de modelo:** é a declaração (ou função) para determinar quão bem um determinado padrão (modelos e seus parâmetros) atende aos objetivos do processo de KDD, ou seja, é avaliada a precisão com que o algoritmo realiza o trabalho de mineração.
- **O modelo de pesquisa:** a função principal deste modelo é encontrar os parâmetros e modelos do método utilizado para otimizar os critérios de avaliação, obtendo resultados precisos.

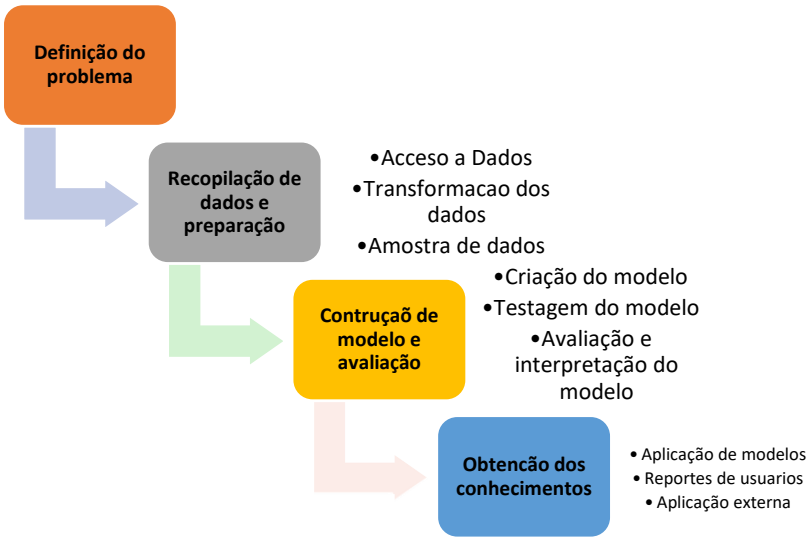


Figura 2.5Processo de mineração de dados.(ORACLE, 2008)

A Figura 2.5 mostra o processo de iteração de um projeto de mineração de dados, que é composto por diferentes fases definidas por Oracle (2008) como:

- **Definição do problema:** Esta fase é o início do processo de mineração de dados e seu foco principal é entender quais são os objetivos e requerimento do projeto a ser desenvolvido.
- **Recopilação de dados e preparação:** a primeira parte implica a compreensão,

recopilação e exploração dos dados. Este processo permite saber se os dados correspondem com os objetivos e requerimentos do projeto. A segunda parte é a preparação dos dados, o que inclui a criação de tabelas, construção dos modelos e preparação dos dados.

- **Construção de modelos e avaliação**: nesta fase devem ser aplicadas diversas técnicas de modelagem e calibração dos parâmetros para obter os valores mais apropriados. Este passo implica a seleção do algoritmo (s) e modelo mais adaptados aos objetivos do projeto.
- **Obtenção de conhecimentos:** é a finalidade da mineração de dados. Nesta fase, os dados que foram processados nos passos anteriores são transformados em informações que poder ser utilizadas para os processos de KDD seguintes.

2.1.10 Aprendizado não supervisionado

A aprendizagem não supervisionada ou não dirigida é um tipo de aprendizagem no qual não existe distinção entre os atributos dependentes e independentes, não existindo resultados anteriores para guiar na construção do modelo. A aprendizagem não supervisionada pode ser utilizada para fins descritivos, assim como também para fazer previsões. Este tipo de aprendizagem é aplicado para um conjunto de dados, onde se precisa descobrir suas relações, agrupamentos e associações, entre outros tipos de informações implícitas (HAN, KAMBER e PEI, 2012).

2.1.11 Agrupamento (Clustering)

Os agrupamentos (clustering) são elementos essenciais a vida dos seres humanos. Desde os primórdios de nossa existência, tentamos agrupar e classificar tudo o que nos rodeia, por exemplo: animais (mamíferos, herbívoros, carnívoros, répteis, etc.), montanha (cordilheira, serras, montes, etc.), terras (África, América, Ásia, Antártida, etc.), que são determinados segundo as características semelhantes que compartilham com cada um de seus componentes ou indivíduos. No aprendizado de máquina, a função principal da análise de agrupamentos é o descobrimento de grupos e padrões com características similares, dentro do conjunto de dados de entrada(ROKACH, 2009).

A análise de agrupamentos organiza os dados mediante a abstração da estrutura subjacente, tanto de agrupamentos de indivíduos ou de hierarquia de grupos. A representação obtida pode conformar grupos de acordo com a ideias preconcebidas ou pode sugerir novos

experimentos (JAIN e DUBES, 1998). Segundo Rokach (2009), as técnicas de clustering agrupam as instâncias de dados em subconjuntos de tal forma que as instâncias semelhantes são agrupadas no mesmo grupo, enquanto que as instâncias diferentes são ordenadas em grupos dessemelhantes. As instâncias são assim organizadas em uma representação eficiente que caracteriza a população o conjunto de dados que está sendo estudado.

Os métodos de agrupamentos são um tipo de aprendizagem computacional por descobrimento muito similar à indução (BERZAL, 2005). É uma aprendizagem não supervisionada, que permite a exploração de dados científicos, de bioinformática, recuperação de informação, entre outras funções. Os agrupamentos também podem ser utilizados para encontrar divergências entre os diferentes grupos que estão sendo analisados (RODRÍGUEZ; CUADRADO; SICILIA, 2007). No final, o objetivo das técnicas de agrupamentos é descobrir um novo conjunto de categorias, os novos grupos são de interesse em si mesmos e sua avaliação é intrínseca (ROKACH, 2009).

O processo de agrupamento é composto pelos seguintes passos, (JAIN e DUBES, 1998):

- **Coletar dos dados:** inclui a extração cuidadosa de dados relevantes a partir das fontes de dados.
- **Triagem inicial**: refere-se à manipulação dos dados após a sua extração, que inclui a limpeza, eliminação de erros e dos itens que serão processados.
- **Representação:** inclui a preparação adequada dos dados, a fim de torná-los adequados para o algoritmo de agrupamento. Aqui são escolhidas a medidas de similaridade, assim como são analisadas as características as dimensionalidades.
- **Tendências de agrupamentos**: verifica se os dados utilizados têm tendência natural a se agrupar ou não.
- **Validação:** muitas vezes são baseadas em técnicas manuais e visuais, isto vai depender da quantidade de dados processados.
- **Interpretação:** esta fase inclui a combinação dos resultados de agrupamento com outros estudos ou referências, a fim de construir as conclusões e sugerir as recomendações.

2.1.12 Classificação dos métodos de agrupamentos.

Existem diferentes tipos de algoritmo de agrupamentos na literatura especializada. A

escolha do algoritmo depende do tipo e a distribuição dos dados, assim como o propósito particular da aplicação desenvolvida. Os algoritmos podem ser classificados nas seguintes categorias (HAN, KAMBER e PEI, 2012):

- **Método de particionamento**: recebe um banco de dados com *n* objetos ou um conjunto de dados. O método de particionamento constrói *k* partições dos dados, em que cada partição representa um grupo $k \leq n$, que deve satisfazer os seguintes requerimentos - 1) cada grupo contém pelo menos um elemento; 2) cada elemento pertence exatamente a um só grupo. Para este método, os algoritmos mais populares são *k-means* e *k-medoids*.

- **Método Hierárquico:** cria uma decomposição hierárquica a partir de um conjunto de dados ou banco de dados. O método hierárquico pode ser classificado como aglomerativo ou divisionista, baseado em como a decomposição hierárquica está formada. A abordagem aglomerativa, também chamada *bottom-up,* inicia-se com cada objeto formando um grupo separado. Combina sucessivamente os objetos próximos um para outro, até que todos os grupos sejam fundidos em um (o nível mais alto de hierarquia), ou até se alcançar a condição de parada. A abordagem divisionista, também chamada *top-down*, começa com todos os objetos no mesmo cluster ou grupo, sendo que a cada iteração o grupo é dividido em grupos menores, até obter os agrupamentos finais ou a condição de parada. Como exemplo de algoritmos aplicados a esta abordagem, pode-se citar: Clustering Using Representatives (CURE), Chamalean, AGNES (*AGglomerative NESting*) e DIANA (*Divisive Analysis*).

- **Método baseado em densidade**: a maioria dos métodos de partição de objetos é baseada na distância entre objetos. Tais métodos só podem encontrar clusters em forma esférica e têm dificuldade para encontrar clusteres em forma arbitrária. Outros métodos de clusteres são desenvolvidos com base na noção de densidade. Nesta categoria, encontramos os seguintes algoritmos: DBSCAN (Density-*Based Clustering Method Based on Connected Regions with Sufficiently High Density*) e OPTIC (*Ordening Points to Identify the Clustering Structure*), entre outros.

- **Métodos baseado em grade (grid):** este método quantifica o espaço entre os objetos de um finito número de células, que formam a estrutura de grade. Todas as operações do cluster são executadas na estrutura da grade. A principal vantagem desta

abordagem é sua alta velocidade de processamento, que geralmente é independente do número de dados, mas dependente do número de células em cada dimensão do espaço quantificado. Dentre estas abordagens, podemos encontrar: Wave Cluster (*Clustering Using Wavelet Transformation*) e CLIQUE (*Clustering High-Dimensional Space*)

- **Métodos baseados em modelos:** presumem a adequação de um modelo para cada um dos agrupamentos e encontra o melhor ajuste dos dados para o modelo dado. Um algoritmo baseado em modelo pode localizar aglomerados através da construção de uma função de densidade que reflete a distribuição espacial dos pontos dados. Também pode determinar de maneira automática o número de grupos baseado em estatística padrão, tendo em conta o "ruído" e os dados atípicos, produzindo um método robusto de detecção. Nesta categoria encontramos os algoritmos Expectation Maximization (EM).

Nos métodos de agrupamento, existem diferente tipos de algoritmos que podem ser usados segundo as características e as finalidades dos projetos. Cada um destes algoritmos utiliza diferentes metodologias - tanto matemáticas como estatísticas - para obter os resultados finais e as conclusões. Dentre os diferentes algoritmos que podem ser usados encontramos os seguintes:

2.1.13 Métodos de agrupamentos

2.1.13.1 Algoritmo K-means

O algoritmo pertence à categoria de métodos de particionamento (Figura 2.6). Está baseado na técnica de centroide para determinar a quantidade de agrupamentos ou cluster. O funcionamento do algoritmo é o seguinte: é dado um conjunto de dados D, o qual contém n objetos em um espaço euclidiano. O algoritmo distribui os objetos de D em k agrupamentos (cluster), C1..., Ck, onde $C \subset D$ e $Ci \cap Cj = \emptyset$ para $(1 \leq i, j \leq k)$. Uma função objetivo é utilizada para avaliar a qualidade de particionamento, para comprovar que os objetos dentro de um grupo (cluster) são semelhantes entre si, mas diferentes dos objetos em outro grupo (cluster). A função objetivo procura alta similaridade intragrupo e baixa similaridade intergrupo. As técnicas de particionamento baseadas em centroide usa o centro do grupo C1 para representar o grupo (cluster). O conceito de centroide pode ser definido como o ponto central do grupo (cluster). O centro poder ser determinado por meio de diferentes formas como média ou medoides dos objetos (pontos) atribuídos ao grupo (cluster) (HAN, KAMBER e PEI, 2012).

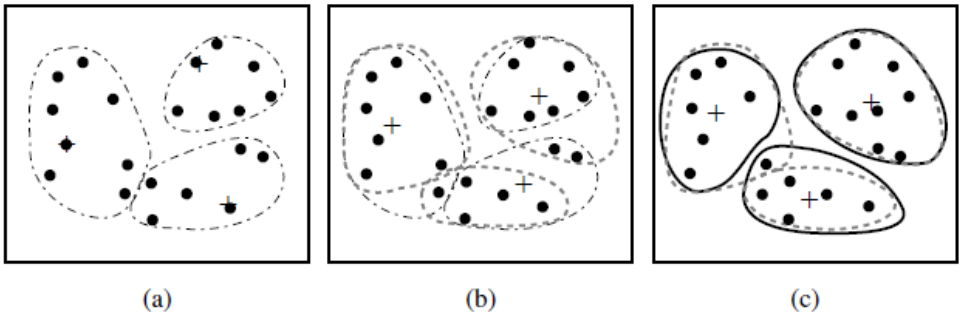


Figura 2.6 Agrupamentos pelo algoritmo K-means. (HAN; KAMBER; PEI, 2012) p.453

2.1.13.2 Algoritmo Hierárquico

Rokach,(2009) indica que esses métodos constroem os cluster, dividindo recursivamente as instâncias de cima para baixo ou de baixo para cima. Estes métodos podem ser subdivididos da seguinte forma: (ROKACH, 2010)

- **Aglomerativo**: cada instância representa um próprio cluster. Em seguida os cluster são combinados até obter o dendrograma final, que é a quantidade de cluster desejados.
- **Divisível**: todas as instâncias pertencem inicialmente a um cluster. Em seguida, o cluster é dividido em subcluster, que são sucessivamente divididos em seus próprios subclusters. Este processo continua até que a estrutura de cluster desejada seja obtida.

Com os resultados dos métodos hierárquicos temos um dendrograma (Figura 2.7), representado por agrupamentos dos objetos e níveis de similaridade no qual as instâncias pertencem. Um agrupamento dos objetos dados é obtido cortando o dendrograma no nível de similaridade desejado (ROKACH, 2009).

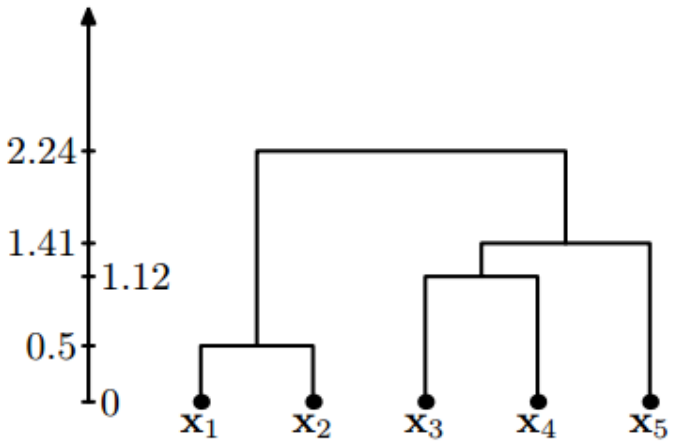


Figura 2.7 :Dendrograma resultado do algoritmo hierárquico (GUOJUN, GAN; CHAOQUN, MA; JIANHONG,

2007) p.142

2.1.13.3 Self-Organization Map (SOM)

Este tipo de algoritmo representa cada cluster como um neurônio ou protótipos. As entradas de dados são representas pelos neurônios, os quais são conectados aos neurônios protótipos. Cada conexão entre os neurônios tem um peso, o qual vai-se adaptando durante o processo de aprendizado(ROKACH, 2009).

Este algoritmo constrói uma rede de camada única. O processo de aprendizado acontece pela abordagem "vencedor leva tudo":

- Os neurônios protótipos competem pela instância que está entrando na rede. O neurônio vencedor é o neurônio cujo vetor de peso é mais próximo à instância que está sendo atualmente representada na rede.
- O vencedor e seus vizinhos aprendem por ter seus pesos ajustados.

O algoritmo SOM é utilizado com sucesso para a obtenção de agrupamentos em dados, para a visualização e o reconhecimento de voz (ROKACH, 2009).

2.1.14 Métricas de avaliação de cluster

As métricas de validação de cluster são índices independentes dos algoritmos que permitem a avaliação da qualidade dos particionamentos para cada conjunto de dados. Tal que as métricas devem ser capazes de selecionar, para cada algoritmo que está sendo avaliado, o conjunto ideal de parâmetros de entradas ou agrupamentos para um conjunto específico de dados (HALKIDI; VAZIRGIANNIS, 2001).

Em geral, as métricas de avaliação de cluster são baseadas nos critérios fundamentais do clustering: 1) compacidade: os algoritmos procuram manter a variação intracluster pequena 2) Conectividade: este é um conceito mais local, o qual estabelece que os itens de dados vizinhos devem compartilhar o mesmo grupo. 3) Separação: este critério procura a maior ou melhor separação intercluster, grupos separados entre si. Os diferentes índices de avalição de clustering se especializam em uma ou em vários destes critérios para avaliar os agrupamentos gerados pelos algoritmos de clustering (HALKIDI; VAZIRGIANNIS, 2001) (HANDL; KNOWLES; KELL, 2005).

As métricas de avalição de clustering são divididas em dois tipos, dependendo do enfoque e cenários de aplicação dos distintos experimentos. São elas:

- **Avaliação externa**: incluem todos os métodos que avaliam um resultado de agrupamento com base no conhecimento dos rótulos de classe corretos. Estas métricas são úteis para permitir uma avalição inteiramente objetiva e comparação de algoritmos de agrupamentos em dados, para os quais os rótulos de classe são conhecidos por corresponder a estruturas de agrupamentos verdadeiras (HANDL; KNOWLES; KELL, 2005).
- **Avaliação interna:** quando não existem rótulos, são usadas as métricas de avaliação interna. Estas técnicas baseiam sua estimativa de qualidade nas informações intrínsecas aos dados. Especificamente, eles tentam medir, de modo preciso, quanto um determinado particionamento corresponde à estrutura de cluster natural dos dados. Estas técnicas estão limitadas a partições puras, isto é, partições em que a cada item dado é atribuído um rótulo ou agrupamento (HANDL; KNOWLES; KELL, 2005).

2.1.15 Linguagem e ambientes para o aprendizado de maquina

- **GNU R** É uma linguagem e ambiente para computação estatística e gráficos. É um projeto GNU, semelhante à linguagem S e o ambiente que foi desenvolvido na Bell Laboratories, por Jhon Chambers e Colegas. R pode ser considerado como implementação de S. A linguagem fornece uma grande variedade de modelos estatísticos (modelagem linear e não linear, testes estatísticos clássicos, análises de séries temporais, classificação, agrupamentos, entre outros) e técnicas gráficas, sendo altamente extensível. Um dos pontos fortes de R é a facilidade com que podem ser gerados os modelos estatísticos e a qualidade dos gráficos produzidos pelo programa (FREE SOFTWARE FOUNDATION, 2016; R CORE TEAM, 2016).
- **Weka (*Waikato Environment for Knowledge Analysis*)** É uma suíte de bibliotecas de classes em Java, que implementa os mais avançados algoritmos de aprendizagem automática e mineração de dados. (WITTEN et al., 1999). O projeto Weka foi fundado pelo governo de Nova Zelândia, em 1993. O programa tem como objetivo a construção de softwares para a implementação dos mais avançados e novos algoritmos na área de mineração de dados e aprendizagem de máquina, além de investigar sua aplicação em área crítica do país, como a economia e a ciência (HALL et al., 2009).

O núcleo de Weka é formado por um pacote de classes que pode ser acessado desde a suíte a todas as outras classes em Weka. As bibliotecas de Java estão organizadas em pacotes lógicos ou diretórios, que contêm as coleções das diferentes técnicas de mineração (WITTEN et al., 1999).

Weka inclui algoritmo de mineração de dados que permite desenvolver projetos de regressão, classificação, agrupamento (clustering), regras de associação, entre outras facilidades. A arquitetura permite a mineração de dados sofisticada com uma alta qualidade e confiabilidade. Weka conta com uma interface gráfica de fácil acesso, que permite o pré-processamento de dados e a configuração dos parâmetros para o processamento em diferentes algoritmos. Além disso, a suíte conta com console e diferentes métodos, que permitem aos desenvolvedores adição de classes e outras funcionalidades ao Weka (HALL et al., 2009).

2.2 Trabalhos Relacionados

Nesta seção, serão apresentados os trabalhos que foram utilizados como referência para o desenvolvimento e a construção desta pesquisa.

2.2.1 Trabalhos relacionados a identificação de Genes Housekeeping

Em Rocha, Santo e Pacheco (2015b), foi apresentada uma revisão bibliográfica de genes de referências ou housekeeping, para os estudos de expressão gênicas baseados em RT-qPCR. Na pesquisa, houve uma busca por trabalhos dos últimos 5 anos, em que foram identificados genes bacterianos de referências validados através de RT-qPCR. Neste trabalho, houve o uso de técnicas de mineração de texto, em que foram identificados 145 genes bacterianos, os quais foram testados em candidatos a genes de referências ou HKG. Destes genes, 45 foram validados experimentalmente e sua estabilidade de expressão foram verificadas usando as ferramentas de software geNorm e NormFinder.

A pesquisa de Rocha selecionou uma lista de 19 genes bacterianos que foram validados como genes de referências em duas ou mais pesquisas. Estes genes basicamente pertencem a oito classes funcionais do genoma e foram validados em diferentes condições experimentais (ROCHA; SANTOS; PACHECO, 2015b).

No estudo de Carvalho et al.(2014), houve a procura por candidatos referências ou HKG em diferentes conjuntos de dados de RNA-seq, da *Corynebacterium pseudotuberculosis*. Foram selecionados 19 genes para o estudo de seus perfis de expressão em diferentes condições de estresse e crescimento. Desta primeira lista, 8 destes genes foram selecionados para serem incluídos como candidatos a genes HKG. Nesta pesquisa, foram usadas técnicas de RT-qPCR, para a análise dos níveis de transcrição em duas condições e a ferramenta NormFinder para a análise da estabilidade. Nestes estudos, todos os genes propostos como candidatos a HKG apresentaram baixos valores de *Maximum Fold Change* (MFC) e coeficiente de variação menor a 4%, que foi o limiar estabelecido pela pesquisa.

A pesquisa de De Jorge et al. (2007) propôs uma metodologia para a identificação de genes housekeeping ou referências que poderiam ser usadas na normalização de experimentos de RT-qPCR. Para a pesquisa, foram utilizados dados de mais de 13.000 amostras de microarray de genes humanos a fim de identificar os genes mais estáveis. O estudo foi capaz de identificar 14 genes que mostram maior estabilidade em diferentes tipos de células e em várias condições experimentais. Como forma de avalição de estabilidade dos genes, foi utilizado como métrica o coeficiente de variação, no qual houve a procura de genes com menor valor e o *máximo fold change* com um limar <2. A equipe foi capaz de confirmar os genes HKG identificados em dados de outras espécies de mamíferos.

2.2.2 Trabalhos relacionado com Corynebacterium pseudotuberculosis

No trabalho de Soares et al. (2013a) foi sequenciado e montado o genoma da *Corynebacterium pseudotuberculosis* na linhagem equi Cp258. A bactéria foi isoloda de um cavalo com linfagite ulcerativa. O genoma foi sequenciado com a NGS Solid v3 e verificou-se ter ele um tamanho de 2.314.404 pares de bases (pb), além de 2.088 regiões como codificadores de proteínas. A montagem deste genoma pode ajudar no desenvolvimento de vacinas que poderão auxiliar no combate a doenças causadas pela C. *Pseudotuberculosis.*

A equipe de Ruiz et al.(2011) sequenciaram e montaram o genoma de *Corynebacterium pseudotuberculosis* na linhagem Cp1002, extraído de uma cabra no Brasil. Após seu isolamento, o genoma foi sequenciado, tendo sido utilizada a tecnologia SOLID v3. A linhagem apresenta um total de 2.335.112 pb, com um total de 2.111 genes anotados. Este estudo faz uma comparação sobre a função de virulência em duas linhagens de *Corynebacterium pseudotuberculosis* (1002 e cpC231) sendo a última isolada de ovelha na Austrália.

Em Pinto et al.(2014), a *Corynebacterium pseudotuberculosis* foi submetida a 3 condições simulada de estresse (choque ácido, osmótico e térmico), mediante o sequenciamento de transcritos usando a plataforma SOLID^{TM} 3Plus. A hipótese foi identificar novos alvos que potencializam a sobrevivência e replicação do patógeno em ambientes adversos. Como resultado da pesquisa foram identificados 474 genes diferencialmente expressos nas 3 condições. Assim, como foram identificados genes importantes para o processo de infecção, assim como os envolvidos na virulência, defesa contra os estresses oxidativos, adesão e regulação. Os dados obtidos por esta pesquisa de Pinto serão usados como bases para o trabalho que serão desenvolvidos neste estudo.

2.2.3 Trabalhos de relacionados à identificação de housekeeping com Técnicas de AM

Em Dong et al.(2011), é apresentado um método para a predição de genes housekeeping usando análises de transformada de Fourier, modificando uma série de dados temporais de expressão gênica do ciclo celular de Hela, no espectros de Fourier. Foi projetado um modelo de classificação efetivo para discriminar entre HKGS e não HKG, usando algoritmo de aprendizado supervisionado Máquina de Vectores de Suporte (SVM). Este algoritmo conseguiu extrair caraterísticas significativas do espetro, permitindo a identificação dos padrões gênicos específicos de HKGs. Ao utilizar este método, houve a identificação de 510 HKG em humanos. Esta metodologia foi validada com dois conjuntos de dados de outros tecidos, o que demonstrou a eficiência modelo na identificação de HKGs.

No trabalho de De Ferrari e Aitken,(2006) é apresentado um método de classificação de HKG baseado em dados de caraterísticas físicas e funcionais, usando como atributos o comprimento do éxon e a medida de compacidade da cromática de genes já disponíveis em bancos de dados, com o algoritmo de classificação Naive Bayes. Este classificador obteve uma taxa de sucesso de 97% na classificação de HKGs humanos.

A maioria dos trabalhos que foram desenvolvidos para a identificação de HKG exploram as técnicas de aprendizado de máquina de classificação e sumarização, que são técnicas predicativas, sendo que a técnicas descritivas são pouco utilizadas. A proposta deste trabalho de pesquisa, pretende explorar a área de descrição de dados, por meio de técnicas de agrupamentos, para propor e desenvolver uma metodologia de identificação de genes housekeeping, através dos perfis de expressão e as distâncias apresentadas entres os diferentes genes nos conjuntos de dados usados.

3. PROPOSTA E ASPECTOS DA ABORDAGEM

3.1 Visão geral

A abordagem proposta está dividida em três grandes passos (Figura 3.1). Pré-clustering, no qual o conjunto de dados é avaliado para verificar se existem agrupamento significativos, que permitam a aplicação da abordagem. Neste passo, os dados são pré-processados para serem utilizados. São usadas também métricas de avalição interna e estabilidade, a fim de verificar os agrupamentos e qual a melhor divisão de grupos; Clustering, nesta etapa são aplicados os algoritmos de agrupamentos nos dados, com o objetivo de definir agrupamentos com o mesmo perfil de expressão. Além disso, são comparadas as similaridades entre os agrupamentos formados; Pós-Clustering, nesta etapa são identificados os candidatos a HKG presentes no conjunto de dados, a ideia é poder obter estes genes baseados em uma lista de HKG já definida e validada como referência. Estabelecidos os dados, as matrizes de distâncias são calculadas para cada condição de estresse. Após essa fase, há os cortes para obter os possíveis candidatos e, finalmente, a lista é filtrada e avaliada por meio de diferentes métricas, para obter os genes candidatos finais.

3.2 Pre-Clustering

3.2.1 Dados de RNA-seq

Na abordagem proposta, são usados dados expressão de RNA-seq, porque fornecem várias vantagens com relação a outras tecnologias como microarray, como pouco ruído, menos quantidade de bias e um custo menor para sua obtenção. O estudo e análise dos dados de expressão de RNA-seq fornecem a possibilidades de compreender e descobrir novos genes, promotores, isomorfos e outro processos biológicos dentro de um determinado organismo (ZYPRYCH-WALCZAK et al., 2015).

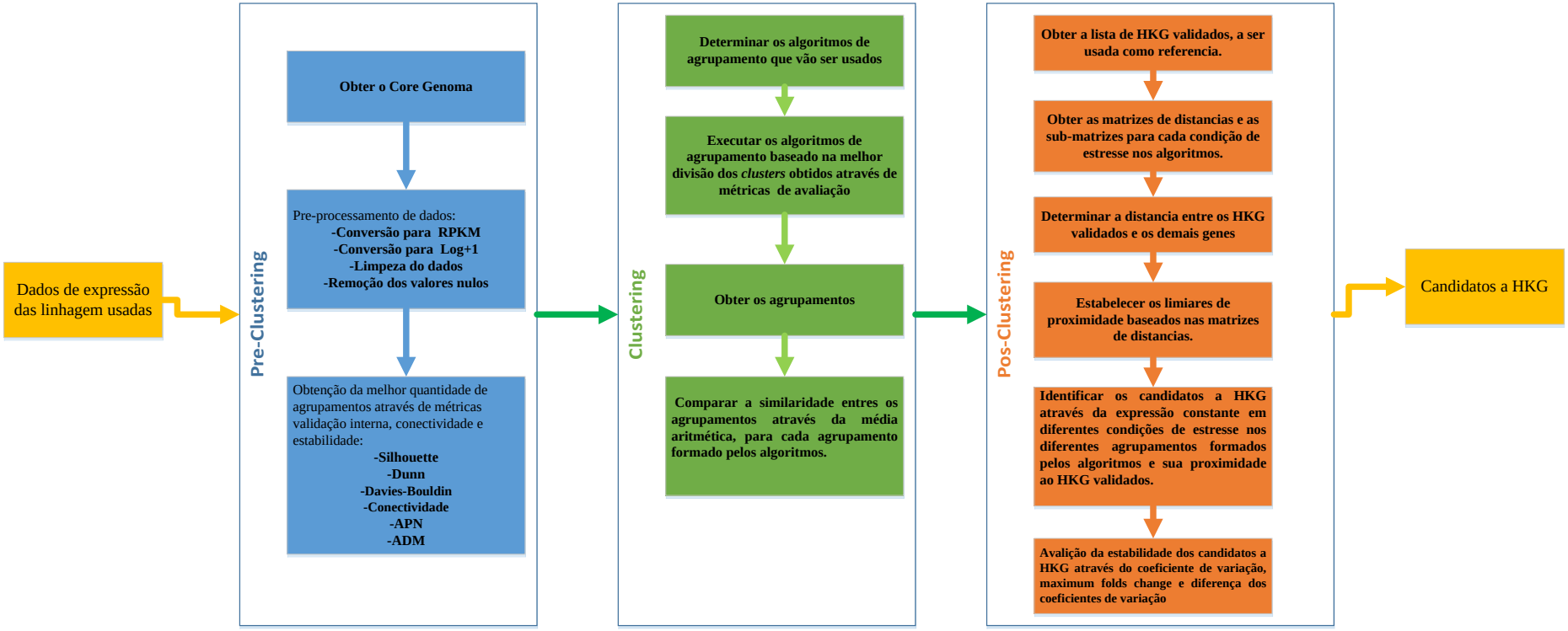


Figura 3.1: Esquema da abordagem para a identificação de genes candidatos a Housekeeping

3.2.2 *Core genoma*

O *core genoma* é o conjunto de genes compartilhado por todas as estirpes de uma espécie bacteriana (WEIGEL, 2014). Tettelin et al. (2005) descobriu que a maioria dos genes presente no *core genoma* da bactéria *S. agalactiae,* pertence ao grupo dos genes housekeeping, os quais têm funções essenciais dentro do genoma, entende-se que estes agrupamentos podem ser descobertos através do uso de técnicas de agrupamentos, o que possibilita a identificação dos HKG. Sabe-se que os HKG são menos propensos a substituição por transferência de genes horizontais do que os restos dos genes do genoma, mas existem exceções conhecidas. Além disso, em Lercher, Urrutia, Hurst (2002), foi descoberto que os genes housekeeping expressos na maioria dos tecidos em humanos mostram fortes agrupamentos, assim como altas taxa de expressão, situação que pode ser similar em organismo procarióticos, como as bactérias.

A identificação *core genoma* permite aumentar a precisão na identificação dos candidatos a genes housekeeping nos conjuntos de dados, considerando que estes são essenciais a sobrevivência da bactéria. Além disso, novos candidatos a HKG podem ser identificados, com base nos agrupamentos que estes formam dentro do genoma.

Na abordagem deste trabalho, a obtenção do core genoma é feito por meio da ferramenta de *software Pan-Genomes Analysis Pipeline* (PGAP)(ZHAO et al., 2012), na qual são analisados os genomas finalizados - dos organismos estudados - que estão depositados nos banco de dados. A ferramenta permite configurar a cobertura e identidade que estamos procurando no pan-genoma para obter o core genoma.

3.2.3 Normalização para RPKM (reads per kilobase per million reads)

As contagens de leituras em bruto não são suficientes para comparar os níveis de expressão entre as amostras, pois esses valores são afetados por fatores como o comprimento de transcrição, o número total de leituras e os vieses de sequenciamentos.

Os procedimentos de normalização tentam facilitar e explicar as comparações precisas entre os grupos de amostra de expressão. Entre os procedimentos de normalização de expressão gênica e transcriptoma mais utilizados temos o RPKM, definido pela equação:

$$RPKM_{gij} = \frac{10^9 . Y_{gij}}{C_{ij} . L_g} \qquad (1)$$

Onde $C_{ij} = \sum_{g=1}^{G} Y_{gij}$ é o total de contagem para o ij-jessima amostra, Lg é o comprimento para o gene e g em número de bases.

3.2.4 Pré-processamento dos dados

O pré-processamento dos dados é uma das tarefas mais importantes para obter um conjunto de dados de boa qualidade, que serão usados nos algoritmos. Este é um dos primeiros processos que devem ser realizados com os dados e é o que toma mais tempo. Existem diferentes técnicas de pré-processamento, como a remoção de ruídos em relação às inconsistências dos dados, a integração dos dados, a redução do tamanho, eliminação de dados nulos, a transformação dos dados, entre outras. Estas técnicas não são exclusivas e podem ser combinadas entre si, a fim de obter um conjunto com menos viés possível (HAN; KAMBER; PEI, 2012). Na abordagem, são realizados basicamente dois processos de pré-processamento, são eles:

- **Limpeza dos dados:** tenta preencher valores que faltam, suavizar o ruído, eliminar ou tratar valores nulos, identificação de outliers e corrigir inconsistência nos dados. Geralmente, este é um processo iterativo de dois passos, no qual temos a detecção de discrepância e o tratamento dos dados (HAN; KAMBER; PEI, 2012).
- **Transformação dos dados:** implica converter os dados em forma adequada para a mineração (HAN; KAMBER; PEI, 2012). Uma das técnicas utilizadas para a transformação de dados de expressão é os logaritmos log (), que permite transformar a distribuição de dados altamente distorcida em distribuições menos distorcidas, permitindo obter uma distribuição mais aproximada à normal. Isto torna os dados mais interpretáveis e acessíveis para trabalhar. Para evitar a influência do valor zero no conjunto de dados é usado o log (x +1), onde x são as instâncias dos dados (HILL, 2012).

3.2.5 Métricas a avaliação interna

Para a avaliação da compacidade e separação dos conjuntos de dados a serem utilizados, são usadas as métricas *Silhouette index*, Dunn index e Davies-Bouldin index. Estas métricas permitem obter a quantidade precisa de agrupamentos para o conjunto de dados segundo o algoritmo usado.

3.2.5.1 Índice Silhouette

Este índice valida o desempenho do agrupamento com base na diferença em pares das distâncias intracluster e intercluster. Além disso, o número preciso é determinado pela maximização do valor deste índice (LIU et al., 2010).

O índice Silhuete é definido como (BOLSHAKOVA; AZUAJE, 2003):

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \qquad (2)$$

Onde *a (i)* é a média da distância entre o ith elemento e todos os demais elementos em xj; "max" é o máximo operador, e *b (i)* é a mínima distância média entre o ith elemento e todo os elementos do cluster em $X_k(k = 1,....c; k \neq j)$. Os valores do índice ficam no intervalo $-1 \leq s(i) \leq 1$, sendo o maior valor o valor ótimo para os agrupamentos (BOLSHAKOVA; AZUAJE, 2003).

3.2.5.2 Índice Dunn

Este índice identifica os conjuntos de agrupamentos (cluster) que são compactos e com boa separação. Para cada partição $U \leftrightarrow X : X_1 \cup ...X_i...X_c$ onde X*i* representa o *i*th elemento do agrupamento para esta partição, o Índice de Dunn, D, é definido como(BOLSHAKOVA; AZUAJE, 2003):

$$D = \min_{1 \leq i \leq c} \left\{ \min_{\substack{1 \leq j \leq c \\ j \neq i}} \left\{ \frac{\delta(Xi, Xj)}{\max_{1 \leq k \leq c}\{\Delta(Xk)\}} \right\} \right\} \qquad (3)$$

Onde $\delta(Xi, Xj)$ define a distância entre os cluster Xi e Xj (distância intercluster); $\Delta(X_k)$ representa a distância intracluster do cluster X_k, e *c* é o número de cluster na partição U. O objetivo desta métrica é maximizar a distância intercluster e minimizar a distância intracluster. O maior valor de D corresponde à melhor quantidade de cluster (BOLSHAKOVA; AZUAJE, 2003).

3.2.5.3 Índice de Davies-Bouldin

O objetivo deste índice é identificar os conjuntos de agrupamentos (cluster) que são compactos e com boa separação. O índice DB está definido como (BOLSHAKOVA; AZUAJE, 2003):

$$DB(U) = \frac{1}{c}\sum_{i=1}^{c} max_{i \neq j}\left\{\frac{\Delta(Xi) + \Delta(Xj)}{\delta(Xi, Xj)}\right\} \quad (4)$$

Onde $\delta(Xi, Xj)$ define a distância entre os cluster Xi e Xj (distância intercluster); $\Delta(Xi)$ e $\Delta(Xj)$ representa a distância intracluster do cluster Xk, e c é o número de cluster na partição U. O alvo desta métrica é maximizar a distância intercluster e minimizar a distância intracluster. O menor valor no DB corresponde aos agrupamentos compactos e com centros longe um do outro (BOLSHAKOVA; AZUAJE, 2003).

3.2.6 Métricas de avaliação da conectividade

3.2.6.1 Índice de conectividade

O índice de conectividade é uma métrica que avalia a medida em que os elementos do conjunto de dados são colocados no mesmo cluster com seus vizinhos mais próximos no espaço (HANDL; KNOWLES; KELL, 2005):

Dado N, denotado como o número total de instâncias (filas) em um conjunto de dados e M denotado como o número total de colunas. Define $nn_{i(j)}$ como o jth vizinho mais próximo do elemento i e deixa $x_{i,nn_i(j)}$ como zero se i e j estão no mesmo cluster e $1/j$ ao contrário. Então, para uma partição particular de um agrupamento $\ell = \{C1, ..., C_k\}$ para N elementos em K agrupamentos disjuntos, a conectividade é definida como (BROCK; PIHUR; DATTA, 2008):

$$Conn(\ell) = \sum_{i=1}^{N}\sum_{j=1}^{L} x_{i,nn_i(j)} \quad (5)$$

Onde L é um parâmetro que dá o número de vizinhos, mas próximo a usar. A conectividade toma valores entre zero e infinito, o melhores resultados desta métrica são os valores mínimos, ou seja as instâncias os conjuntos de agrupamentos que apresenta maior conectividade (BROCK; PIHUR; DATTA, 2008).

3.2.7 Métricas de estabilidade

As métricas de estabilidade permitem avaliar a estabilidade ou consistência dos dados, comparando os resultados do conjunto de dados a partir da remoção de cada coluna por vez. Esta métrica permite avaliar a consistência dos dados com uma quantidade cluster definida (DATTA; DATTA, 2003). No modelo, são utilizadas duas métricas de avaliação de estabilidade.

3.2.7.1 Porção das médias que não se superpõem (APN)

Esta métrica computa a média de porção de elementos (genes) que não são colocados no mesmo cluster pelo algoritmo de agrupamento e os dados obtidos pela remoção de uma coluna (expressão gênica) cada vez no tempo. Está definida como (DATTA; DATTA, 2003):

$$V_1(k) = \frac{1}{M1}\sum_{g=1}^{M}\sum_{i=1}^{1}\left(1-\frac{n(C^{g,i}\cap C^{g,0})}{n(C^{g,0})}\right) \quad (6)$$

A métrica é um intervalo de [0,1], no qual o valor perto de zero corresponde ao cluster com mais alta consistência e estabilidade (DATTA; DATTA, 2003).

3.2.7.2 Medida de distâncias média entre médias

Computa a média da distância entre as médias dos coeficientes de expressão (log transformado) para todos os genes que são colocados no mesmos cluster pelo algoritmo de agrupamento usado, para todo o conjunto de dados e para o dado obtido através da remoção de uma coluna (níveis de expressão) uma vez no tempo. Esta métrica está definida como (DATTA; DATTA, 2003):

$$V_2(K) = \frac{1}{M1}\sum_{g=1}^{M}\sum_{i=1}^{1} d(\bar{\bar{X}}\,C^{g,i}, \bar{\bar{X}}\,C^{g,0}) \quad (7)$$

Onde $\bar{\bar{X}}\,C^{g,0}$ denota a média do perfil de expressão para os genes através do cluster $C^{g,0}$ e $\bar{X}\,C^{g,i}$ denota a média do perfil de expressão através do cluster $C^{g,i}$. O menor valor encontrado representa a melhor quantidade de clusters.

3.3 Clustering

3.3.1 Seleção dos algoritmos de agrupamentos

Neste passo, são obtidos os agrupamentos, a partir dos algoritmos usados, que permitem estabelecer os perfis gênicos das linhagens. A seleção dos algoritmos deve ser feita a partir do tipo de dados que está sendo usado, assim como as características específicas. As características e os tipos de dados vão determinar o tipo de metodologia a ser utilizada.

Para obter melhores resultados na metodologia, vários tipos de algoritmos devem ser testados e combinados. É recomendada a utilização de pelo menos dois tipos de algoritmo diferentes, o que garante maior confiabilidade nos dados obtidos pela abordagem.

Nesta pesquisa, houve o uso de três algoritmos de agrupamento que utilizam diferentes metodologias para gerar agrupamentos. Os algoritmos selecionados foram: particionamento (*K-means*), hierárquico (aglomerativo) e rede neural (*Self Organization Map*-SOM)(ANDRITSOS, 2002). Para a seleção destes algoritmos foram usados dez algoritmos de agrupamentos de diferentes tipos e metodologias, sendo que estes três apresentaram o melhor desempenho devido ao tipo de dados e a distribuição que estes apresentaram.

K-Means é um algoritmo que inicia desde uma porção inicial de objetos (genes) e procede interativamente, calculando o centroide (media) dos agrupamentos e reassinando cada objeto para o cluster mais próximo, de acordo com várias métricas de distância, como a distância Euclidiana. As iterações continuam até ficarem sem objetos para serem assinados nos agrupamentos (SI et al., 2014). O algoritmo precisa que seja especificada a quantidade de agrupamentos para a divisão dos dados.

O método hierárquico constrói os agrupamentos por meio de divisão recursiva das instâncias. Na metodologia aglomerativa (de baixo para cima) cada objeto representa um agrupamento. Então, os agrupamentos são sucessivamente combinados até obter a quantidades de agrupamentos desejados (ROKACH; MAIMON, 2010).

Self Organization Map (SOM), este tipo de algoritmo representa cada cluster como um neurônio ou protótipo. As entradas são representadas por neurônios, os quais são aprendidos de forma adaptativa durante o aprendizado. Este algoritmo constrói uma rede única de camada. O processo de aprendizagem ocorre da forma que o “o vencedor leva tudo”.

- Os neurônios protótipos competem pelas instâncias atuais. O vencedor é o neurônio cujo vetor de peso é mais próximo à instância atualmente apresentada.
- O vencedor e seus vizinhos aprendem por ter os pesos ajustados.

O algoritmo SOM é utilizado com sucesso para a quantificação vetorial e pela rapidez do reconhecimento (ROKACH; MAIMON, 2010).

3.3.2 Implementação dos algoritmos de clustering

Para a implementação dos algoritmos de agrupamentos, é estabelecido um consenso entre as diferentes métricas de avaliação a que foi submetido o conjunto de dados, para a seleção do número ótimo de agrupamentos. Para isto, é verificada a melhor solução de agrupamento apresentado por cada métrica e é selecionado o número com maior frequência nas métricas, obtendo só um número para todos os algoritmos. Também são configurados os atributos específicos de cada algoritmo, como as métricas de distâncias para K-means e hierárquico e a quantidade de neurônios para SOM.

Com base nestas métricas e configurações, são obtidos os agrupamentos de algoritmos que vão servir de base para as análises realizadas na abordagem. Nesta pesquisa para a implementação foram usadas as ferramentas de software R e Weka. O primeiro foi utilizado para a avaliação dos dados e o decibrimentos dos agrupamentos, assim como para o analises gráfico dos agrupamento, já weka foi utilizado para o pré-processamento dos dados e a identificação dos agrupamentos nos diferentes conjuntos de dados.

3.3.3 Comparação da similaridade os agrupamentos das diferentes linhagens

Para se comparar a similaridade dos agrupamentos formados entre as linhagens, foi desenvolvido um procedimento heurístico que permitiu a comparação dos agrupamentos com maior similaridade nos diferentes conjuntos de dados, baseado na média aritmética dos dados nos diferentes agrupamentos formados. O procedimento se constitui de três fases:

- **Comparação do perfil de expressão do agrupamento.**

$$\bar{\bar{X}}_i(c_{1i}) \approx \bar{\bar{X}}_j(c_{1j}) \qquad (8)$$

Onde $\bar{\bar{X}}$ é a média aritmética do conjunto, i e j representam as linhagens usadas e

C os agrupamentos definidos pelos algoritmos. Se a média do perfil de expressão $\bar{\bar{X}}$ no C_{1i} conjunto de dados da linhagem A é aproximada à média de expressão de C_{1j} no conjunto de dados da linhagem B do mesmo microrganismo são próximas, os agrupamentos formados pelos algoritmos são considerados como similares.

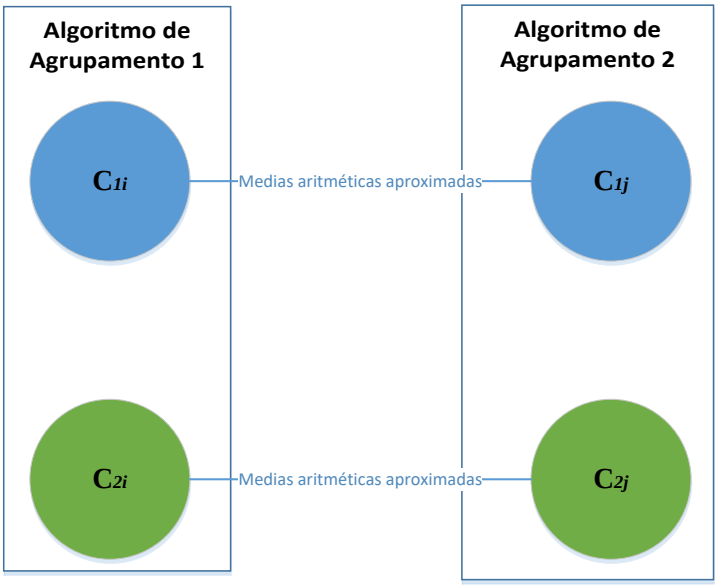


Figura 3.2 Comparação entre agrupamentos de diferentes algoritmos, para estabelecer a similaridades entres os grupos.

- **Tamanho dos agrupamentos**

$$C_{1i} \approx C_{1j} \qquad \text{Equação 1}$$

Se os agrupamentos nos dois conjuntos apresentam um tamanho aproximado, então podem ser considerados como similares.

- **Genes similares**

Se existir o gene $y \in C_{1i}$ no conjunto de dado A e $y \in C_{1j}$ no conjunto de dado B, sendo que y representa um gene qualquer dentro do agrupamento, estes são considerados como similares.

3.4 Pós-Clustering

3.4.1 Lista de genes housekeeping a ser usada como referência.

Para a identificação dos candidatos a HKG, é preciso identificar genes de referência ou housekeeping validados em bancada, por técnicas como RT-qPCR ou outras, e que tenham sido descritos na literatura, para o organismo estudado ou para o domínio biológico ao qual este pertence. Isto é importante, pois segundo Lecher, Urrutia, Hurst(2002), os genes housekeeping mostram fortes agrupamentos em diferentes tecidos. Com base nessas informações, podemos ter a premissa que, se temos um HKG identificado na literatura, é provável que genes ao seu redor sejam HKG.

3.4.2 Cálculo das matrizes de distância.

Dada uma lista de referência de genes validados como HKG, e feitos os cálculos das matrizes de distância de cada uns dos genes do conjunto para os genes HKG de referência. Para obter as matrizes é utiliza-se a distância Euclidiana: esta é comumente usada como métrica de dissimilaridade, assim, a distância euclidiana entre os pontos X e *Y* é calculada como:

$$d_{euc}(x,y) = \left(\sum_{j=1}^{p} (x_j - y_j)^2 \right)^{\frac{1}{2}} \quad \text{Equação 2}$$

A distância Euclidiana foi obtida para cada um dos dados em cada uma das condições de estresse, com base nos valores de expressão gênica.

$$d_{euc} \rightarrow A \subseteq B \quad \text{Equação 3}$$

Onde A é uma linhagem e B é uma condição de estresse específica. B é subconjunto de A em uma condição específica na qual foi submetido o microrganismo.

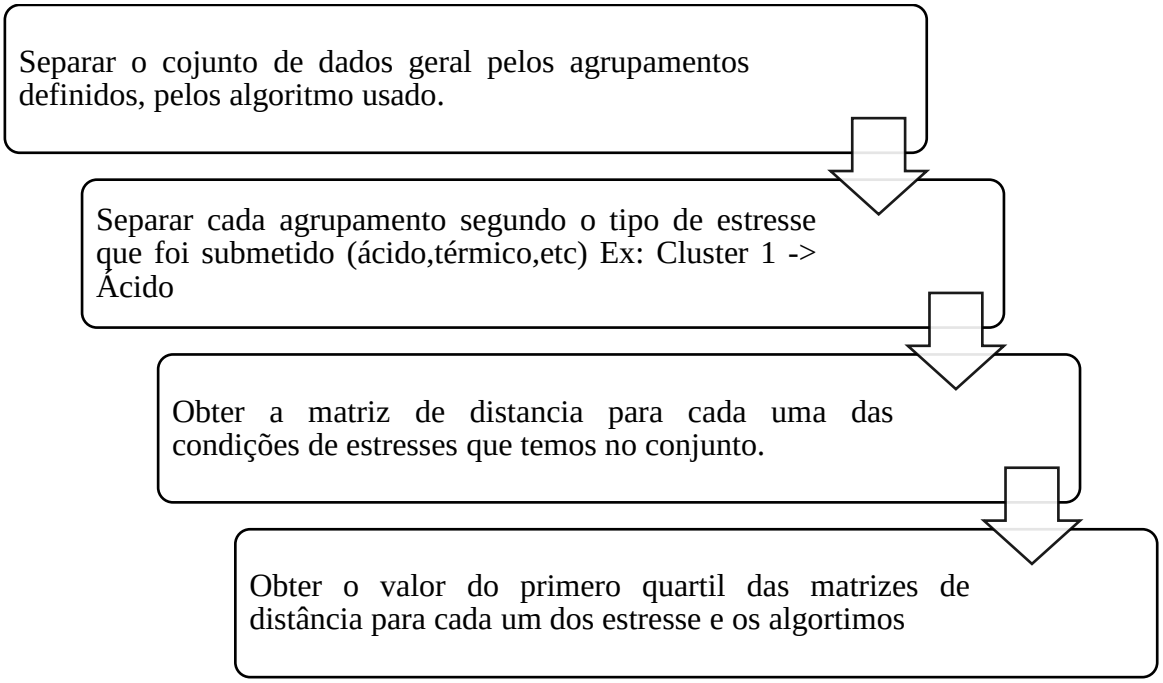


Figura 3.3: Método para a obtenção das matrizes de distância, segundo o tipo de estresse usado e os algoritmos.

Para a obtenção das matrizes, foi desenvolvido o método heurístico (Figura 3.3), o qual permite obter valores de distância segundo o tipo de estresse utilizado e os agrupamentos. Este método gera uma matriz para cada tipo de estresse dentro dos agrupamentos definidos pelos algoritmos usados. Após obter as matrizes, são obtidas as métricas da estatística descritiva (media, quartilhes, mediana, etc.). A partir do valor do primeiro quartil (valor selecionado nesta pesquisa) são obtidas as submatrizes de distâncias que vão permitir identificar os possíveis genes candidatos a HKG.

3.4.3 Criação das submatrizes baseado no limiar de corte.

Para seleção dos genes mais próximos, são criadas submatrizes a partir da matriz euclidiana. Para isto, é utilizado o primeiro quartil (q1) da distância como limiar de corte, já que este apresenta os elementos que são próximos aos genes HKG validados. Então, se $y \leq q_1$ onde o gene *y* é menor que q1 em uma condição de estresse específica. Para isto, é usado o algoritmo 1, que permite a criação das diferentes submatrizes para cada uma das condições de estresse.

Entrada -> Matriz de distancia: d_{euc} [y_1, y_2…y_n]

Saída -> Matriz Quartil1 Mq1 [y_1, y_2...y_n]

/ cálculo do primeiro quartil da matriz de entrada /

$Q1 = (d_{euc} + 1) / 4$

/Criação da matriz de saída /

Para cada y_i em d_{euc} **faça**

Se $y_i \leq Q1$ e**ntão**

Escreve y_i em Mq1 [y_i...y_n]

Fim se

Retorna Mq1

Algoritmo 1: Algoritmo criação da submatrizes de distancias baseado no limiar de corte estabelecido.

3.4.4 Identificação dos possíveis candidatos a genes housekeeping

Para a identificação dos possíveis candidatos a HGK, são analisadas as distâncias (figura 3.4) de cada um dos genes presentes nas submatrizes nas diferentes condições, com relação com os HKGs validados, sendo selecionados como possíveis candidatos os genes que apresentam menor proximidade.

A partir das submatrizes de distância euclidiana, vai ser possível a identificação dos genes que estão mais próximos dos genes HKG validados (lista de referência) e cuja expressão gênica é constante nas diferentes condições. A partir disso, será possível selecionar os candidatos a genes housekeeping. Para que isto ocorra, uma matriz é criada com relação a um gene de referência, que combina as submatrizes de distância nas diferentes condições deestresses que estão sendo estudadas. Nestas novas matrizes, são procurados os genes que apresentam uma expressão constante nas diferentes condições, ou seja, que apresentam proximidade com relação aos genes de referência em todas as condições, com pouca variação de expressão.

Entende-se que se *y* é um gene HKG da lista de referência, em C_{1j}, onde C é um agrupamento em uma linhagem específica, que apresenta uma expressão constante com pouca

variação. Já *A* é um gene que em todas as condições apresenta pouca variação do nível de expressão, e com relação ao (q1) está perto de *y*, então pode-se indicar que *A* é um possível candidato a gene housekeeping para a linhagem em questão (figura 3.5).

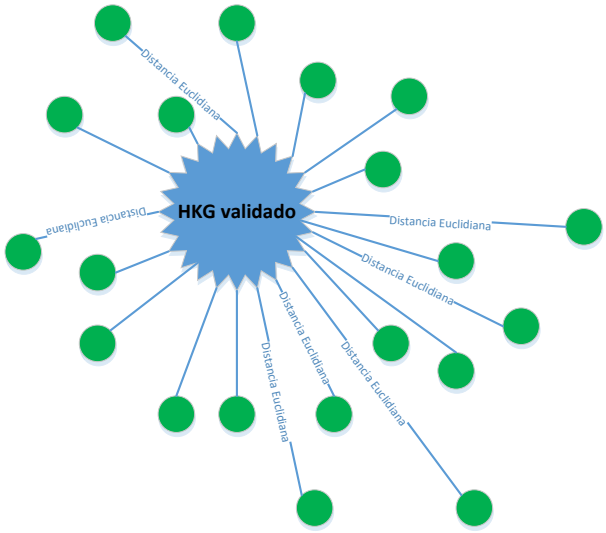


Figura 3.4: Análises da distância entre os HKG validados e os genes da submatrizes, para a seleção dos possíveis candidatos a HKG

Os conjuntos de genes obtidos através desta técnica são comparados, para encontrar os genes comuns entres a diferentes condições figura 3.5, de um conjunto de dado e com um algoritmo de cluster especifico.

.

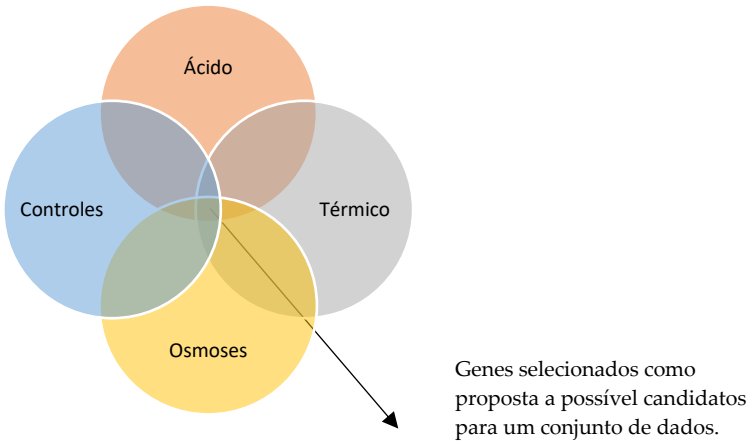


Figura 3.5 Na figura A, temos a representação de um possível candidatos a HKG, baseado na expressão constante e proximidade com relação a um ou vários dos genes validados, nas diferentes subcondições

Os genes obtidos como possíveis candidatos, a partir das submatrizes euclidianas nas

diferentes condições de estresse, do conjunto de dados A, são comparados ao resultado do conjunto B, gerado com o mesmo algoritmo de agrupamento e no agrupamento com um perfil de expressão similar, $y_{1j} \in C_{1j}$ e $y_{1i} \in C_{1i}$. Logo, o gene pode ser considerado como um possível candidato a housekeeping.

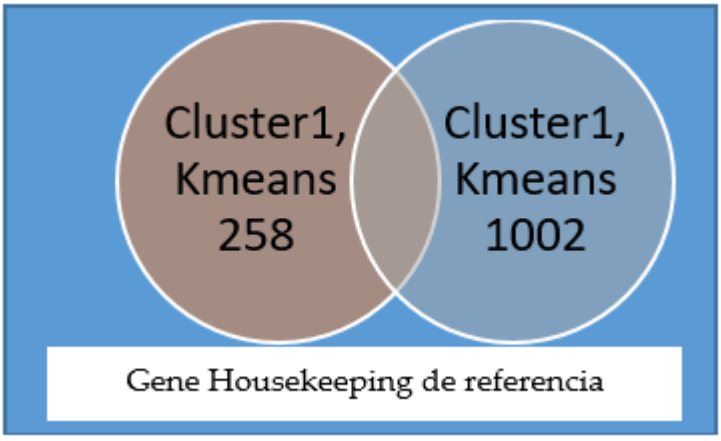


Figura 3.6: Comparação de genes que pertence a agrupamento com perfis aproximado nos diferentes conjuntos de dados com relação ao gene de referência identificado.

3.4.5 Validação dos possíveis candidatos a HKGs

Um gene candidato a housekeeping é definido como um gene com expressão mais estável e pequenas variações de expressão (DE JONGE et al., 2007). Baseado nesta definição, para avaliar os genes que foram identificados como possíveis candidatos, é realizado um teste de estabilidade através do coeficiente de variação (CV). Este teste é usado como métrica estatística para comparação do grau de variação entre os genes, independentemente de sua expressão média (DE JONGE et al., 2007). O CV está definido pela equação:

$$cv = \frac{S}{X} \qquad \text{Equação 4}$$

Onde S é o desvio padrão da expressão dos genes e X é a média da expressão do gene. Esta técnica possibilitou a seleção dos genes com menor variação de expressão, pelo que foi determinado como limiar os CV <10.

Outra métrica para a avaliação dos candidatos a housekeeping foi o *Máximo change fold* (MFC). Esta métrica representa a menor variação na expressão dos HKG (DE JONGE et al., 2007). A MFC é calculada pela equação:

$$MFC = \frac{Max\ y_1}{Min_{y1}}$$ Equação 5

Onde Maxy_1 e Miny_1 presentam o maior e menor valor de expressão dos genes nas diferentes condições do conjunto. Para a seleção dos genes, foi determinado o limiar de <2(DE JONGE et al., 2007).

Entende-se que a expressão de um HKG, situada entre uma linhagem e outra do mesmo microrganismo, o qual foi submetido às mesmas condições, não deve apresentar uma variação tão diferenciada, pois o uso da diferença entre os coeficientes de variação procura a menor variação possível.

Para validar as diferenças de expressão entre linhagens, foi utilizada, de forma heurística, a equação:

$$Dcv = Cv_1 - Cv_2$$ Equação 6

Em que Cv_1 e Cv_2 são os coeficientes de variação para cada gene nas linhagens. Esta métrica permite conservar os genes que apresentam menor variação de coeficientes de expressão entre as linhagens estudadas.

Estas avaliações permitem que sejam identificados como candidatos a genes housekeeping aqueles que apresentam uma expressão estável, com menor coeficiente de variação, como um máximo *fold change* baixo em que a diferença de variação entre linhagem seja baixa.

4. O ESTUDO DE CASO

Neste capítulo, serão apresentados o estudo de caso no qual a abordagem foi aplicada e os resultados obtidos a partir da aplicação.

Como organismo-alvo para o estudo de caso, foi utilizada a *Corynebacterium pseudotuberculosis*, uma bactéria grande-positiva, não metil, pleomórfica e facultativa da ordem de *Actinomycetales*. É um microrganismo intracelular facultativo, que pode proliferar dentro dos macrófago (SOARES et al., 2013a). Esta bactéria é organismo causador de uma doença crônica em gado - ovinos e caprinos (SELIM et al., 2016).

4.1 Obtenção dos dados

O dados foram obtidos a partir da pesquisa de Pinto et al., (2014) que submeteram os dados a diferentes condições de estresse, para identificar a resposta da *Corynebacterium pseudotuberculosis* nestas condições. Para isto, usaram: osmoses conseguido com 2M NaCl; estresse ácido, com suplementação de ácido clorídrico com um pH de 5; estresse térmico induzido por ressuspensão da pastilha em meio BHI pré-aquecido a 50º C e uma condição controle do experimento.

Os dados foram sequenciados na tecnologia SOLID™, utilizando RNase III para a preparação das bibliotecas de amplificação de cDNA, a qual foi produzida por transcrição reversa a partir de adaptadores ligados às extremidades da molécula de RNA, de acordo com o protocolo SOLID™ Total RNA-Seq kit (*Life Tecnhologies™).*

4.2 Pré-clustering

4.2.1 Core genoma

Para a obtenção do core genoma, nesta pesquisa foram utilizadas 38 linhagens da Corynebacterium pseudotuberculosis (12 equinos e 26 ovinos) através do software Pan-Genomes Analysis Pipeline (PGAP)(ZHAO et al., 2012), com uma cobertura de 90% e um e-value de 1E05.

Obtendo um core genoma de 1.285 genes para a CP258 e de 1191 genes para a CP1002. Os genes contidos no core genoma foram comparados com o conjunto das linhagem 258 e 1002 da *Corynebacterium pseudotuberculosis* a fim de obter os genes comuns expressos a partir dos

dados obtidos pelo experimento de Pinto e companheiros (PINTO et al., 2014).

Observando o conjunto de genes pertencentes ao core genoma, para a linhagem 258, 1.191 genes (55% do total do genoma) apresentaram expressão nas diferentes condições de estresse estudadas. Para o genoma de 1002, 1.141 genes (54% do total) estavam expressos (Tabela 4.1).

4.2.2 Pré-processamento de dados

Os dados obtidos a partir do core genoma foram submetidos a um processo de limpeza, em que foram eliminados todos os genes que apresentaram expressão nula, sendo 17 genes (1.4%) do total do core genoma para a CP258 e para a CP1002 12 genes (1.05%) do total do core genoma. Tendo como resultado final um conjunto de dados de 1174 para CP258 e de 1129 pra CP1002 (Tabela 4.1)

Linhagem	A- Total de genes do genoma (instâncias)	B-Genes no core genoma total (instâncias)	C-Genes do core genoma expressos nas linhagens	D- Instâncias eliminadas pelo pré-processamento (genes com valores nulos)	E-Conjunto de dado final
Cp258	2137	1285	1191	17	1174
Cp1002	2100	1191	1141	12	1129

Tabela 4.1: Processamento dos dados para a obtenção do conjunto de dados finais, a partir das linhagens e o core genoma.

Os dados foram normalizados a RPKM, segundo a equação 1, o que permite a comparação entres os dois conjuntos de dados que estão sendo usados nesta pesquisa.

Os dados originais mostravam uma distribuição com concentração para a esquerda (Figuras 4.1. e 4.2). Devido a este fator, os dados foram transformados em logaritmo log (x+1) para obter uma distribuição a mais próxima da normalidade (Figuras 4.3 e 4.4).

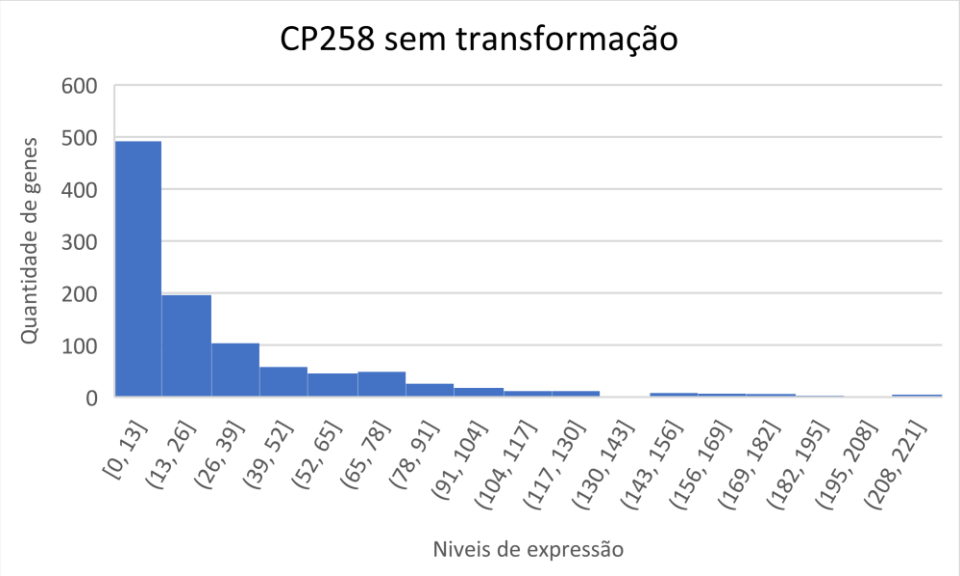


Figura 4.1: Distribuição da Linhagem CP258 antes da transformação de log (x+1). Mostra a distribuição com concentração para a esquerda dos dados

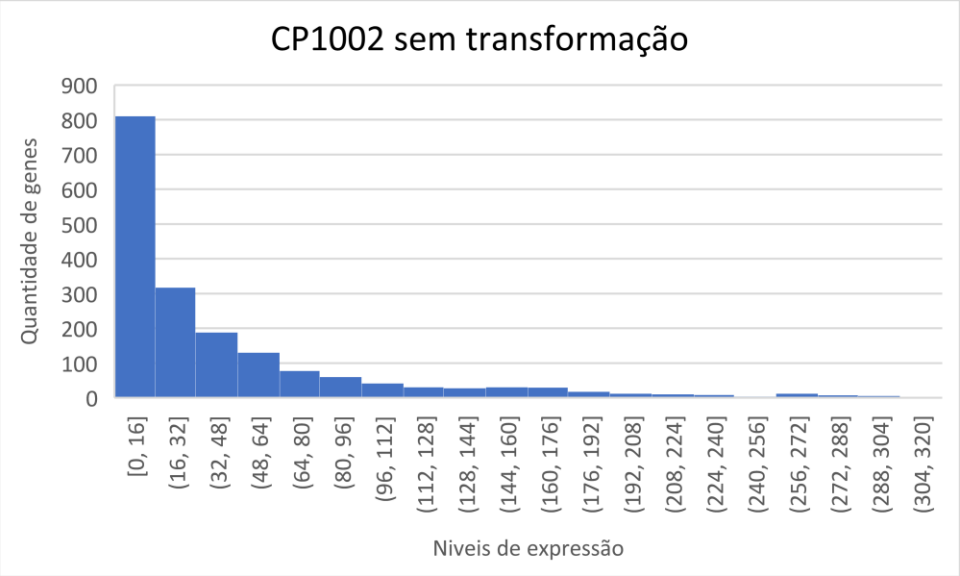


Figura 4.2 Distribuição da Linhagem CP1002 antes da transformação de log (x+1). Mostra a distribuição com concentração para a esquerda dos dados

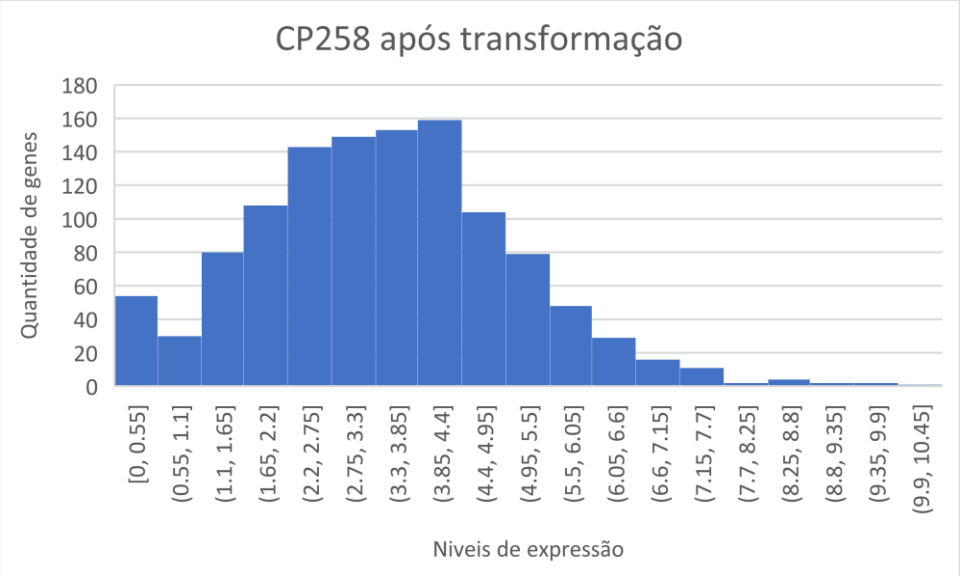


Figura 4.3 Distribuição da Linhagem CP258 após da transformação de log (x+1). Mostra a distribuição mais próxima da normalidade.

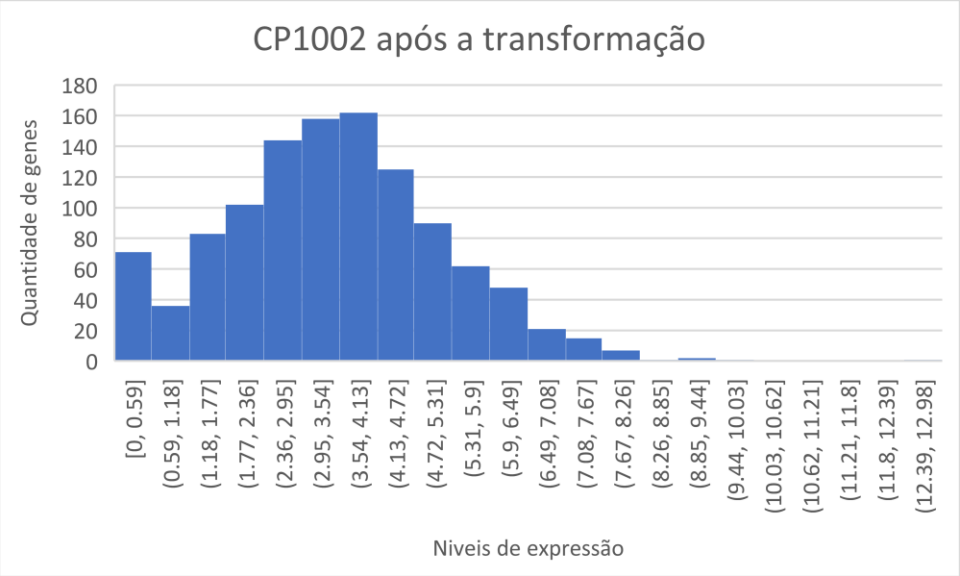


Figura 4.4 Distribuição da Linhagem CP1002 após da transformação de log (x+1). Mostra a distribuição mais próxima da normalidade.

4.2.3 Métrica de avaliação interna

Para a implementação da métricas de avalição interna foram utilizados os pacotes Nbclust e Clvalid da suíte R, sendo que o primeiro implementa 30 métricas de avaliação para determinar o número do cluster. A avaliação dos dados usado estas métricas mostraram que a melhor divisão era 2, além do que nesta pesquisa focamos nas métricas definidas anteriormente na seção 3.2.6. Os pacotes Nbclust e Clvalid implementa as métricas de avalição interna que são o alvo de pesquisa nesta pesquisa. As Figuras 4.5 e 4.6 mostram os resultados do pacote

Nbclust onde foram aplicados 30 índices de avaliação interna, para obter a melhor quantidade de agrupamentos para os dos conjuntos de dados que foram usados, o consenso das métricas determinou que a melhor quantidade de agrupamentos para as linhagens é 2.

```
*******************************************************************
* Among all indices:
* 9 proposed 2 as the best number of clusters
* 7 proposed 3 as the best number of clusters
* 1 proposed 4 as the best number of clusters
* 3 proposed 6 as the best number of clusters
* 1 proposed 11 as the best number of clusters
* 1 proposed 12 as the best number of clusters
* 1 proposed 15 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  2


*******************************************************************
```

Figura 4.5: Saída do Pacote Nbclust para a linhagem CP258, com os 30 índices, apontando a 2 como a o melhor número de agrupamento para este conjunto de dados.

```
*******************************************************************
* Among all indices:
* 8 proposed 2 as the best number of clusters
* 7 proposed 3 as the best number of clusters
* 1 proposed 6 as the best number of clusters
* 2 proposed 9 as the best number of clusters
* 4 proposed 11 as the best number of clusters
* 1 proposed 14 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  2


*******************************************************************
```

Figura 4.6 Saída do Pacote Nbclust para a linhagem CP1002, com os 30 índices, apontando a 2 como a o melhor número de agrupamento para este conjunto de dados.

4.2.3.1 Índice Silhouette

Esta métrica foi implementada com os pacotes Nbclust e Clvalid seguindo a equação 2, e apresentou como resultado que o melhor número de agrupamento para os conjuntos de dados é 2 nos diferentes algoritmos que vão ser utilizados (Figuras 4.7 e 4.8).

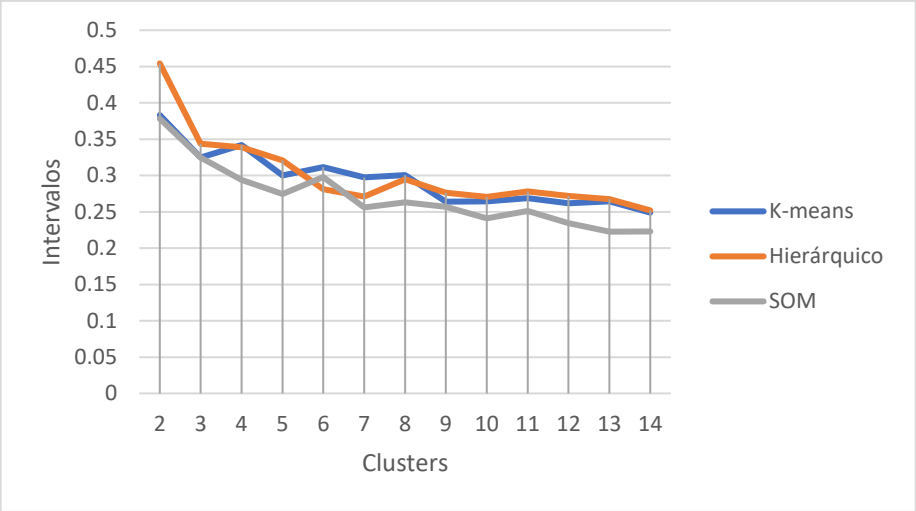


Figura 4.7: Índice Silhouette para a linhagem CP258, apontando que o melhor número de cluster para os diferentes algoritmos é 2

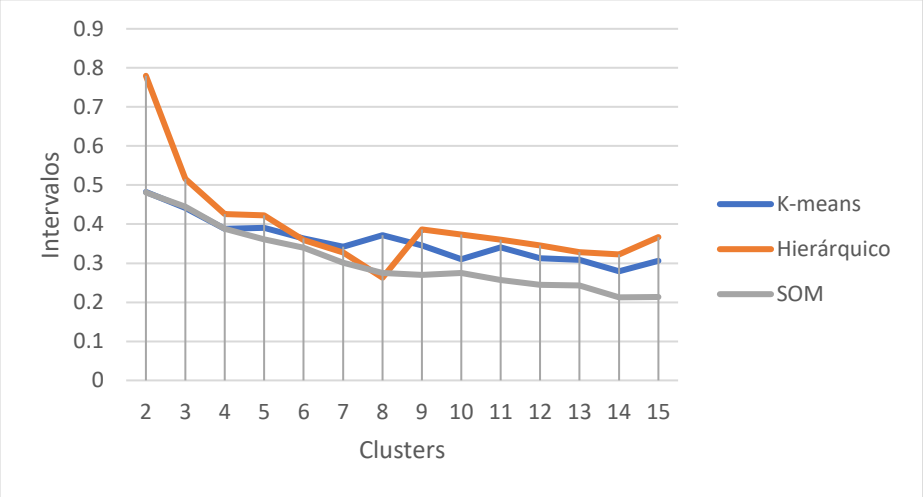


Figura 4.8 Índice Silhouette para a linhagem CP1002, apontando que o melhor número de cluster para os diferentes algoritmos é 2

4.2.3.2 Índice Dunn

Esta métrica foi avaliada através dos pacotes Nbclust e Clvalid em R, seguindo a equação 3, sendo que o melhor resultado para o número de agrupamento foi 2, sendo que a linhagem Cp258 mostrou 3 como o melhor número para o algoritmo SOM. Para a CP1002, os valores são mais heterogêneos, sendo 2 para o algoritmo hierárquico, 14 para o K-means e 15 para SOM, sendo que a diferença de valores não é muito pronunciada no intervalo implementado (Figuras 4.9 e 4.10).

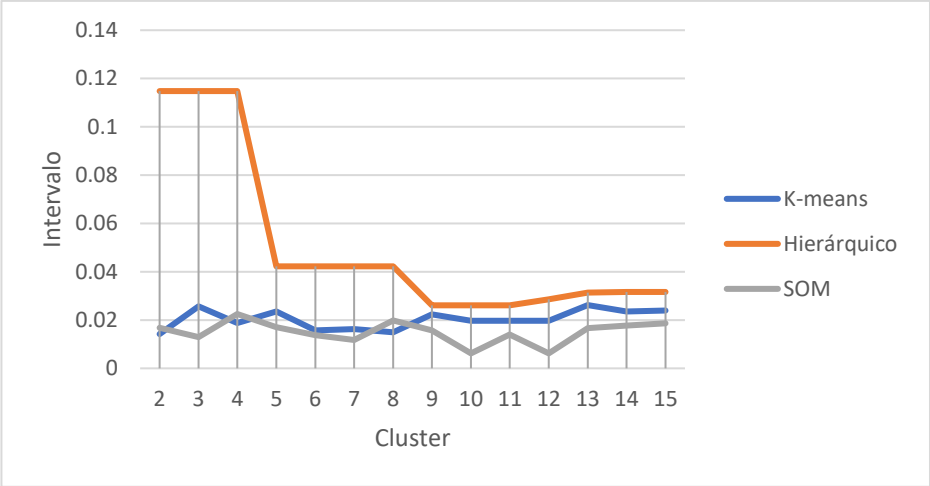


Figura 4.9: Índice Dunn para a linhagem CP258, mostrando 2 como o melhor valor para os algoritmos hierárquico e K-means e 3 como o melhor para SOM

Figura 4.10: Índice Dunn para CP1002, apontando 2 como o melhor número para o algoritmo hierárquico, 14 para os algoritmos K-means e 15 para SOM.

4.2.4 Métrica de conectividade

Com a finalidade de verificar a conectividade dos dados foi utilizada esta métrica com o pacote Clvalid, seguindo a equação 4, sendo que o melhor número de cluster foi 2 para as duas linhagens (Figuras 4.11 e 4.12)

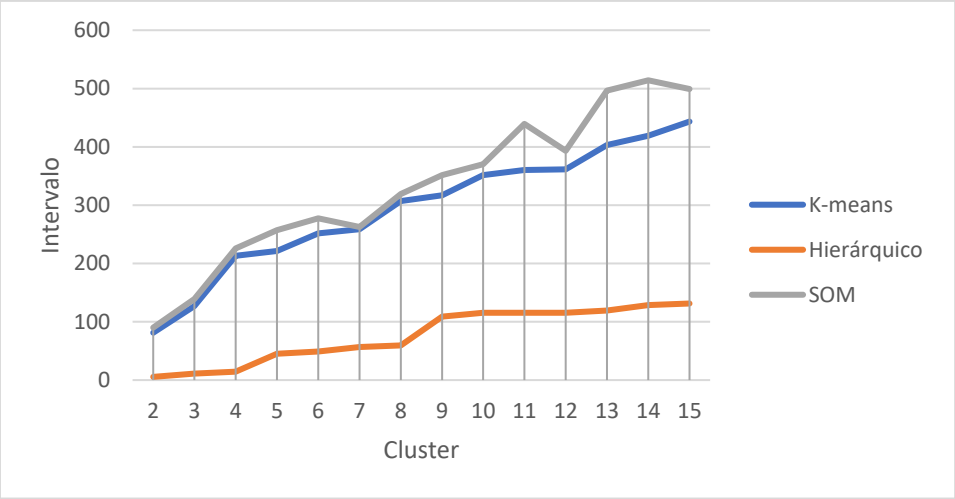


Figura 4.11: Índice de conectividade para a linhagem CP258, indicando 2 como o melhor número de agrupamento para o conjunto

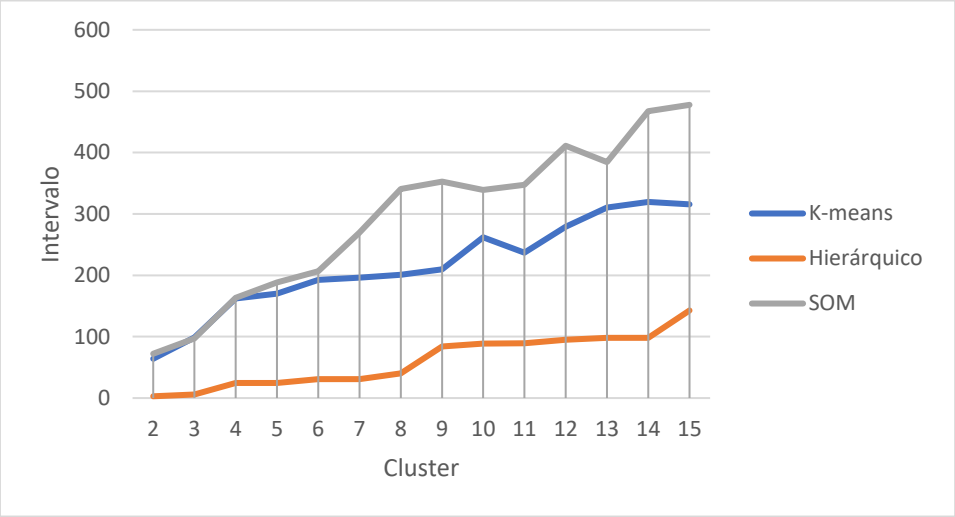


Figura 4.12: Índice de conectividade para a linhagem CP1002, indicando 2 e 3 melhor número de agrupamento para o conjunto de dados

4.2.5 Métrica de estabilidade

4.2.5.1 Porção das médias que não se superpõe (APN)

Esta métrica foi implementada seguindo a equação 6, para obter a estabilidade dos conjuntos de dados baseado nas médias que não se superpõe. Esta métrica indicou como melhor número de cluster 2, para as duas linhagens utilizadas.

Figura 4.13: Métrica APN para a linhagem CP258, indicando que o melhor número de cluster entre 2 e 5, mostrando a melhor estabilidade de dados

Figura 4.14: Métrica APN para a linhagem CP1002 indicando que o melhor número de agrupamento entre 2 e 3 dois mostrando a melhor estabilidade do conjunto

4.2.5.2 Medida de distancias média entre médias

Esta métrica foi implementada baseada na equação 7, para o obter o melhor número de agrupamentos a fim de que o conjunto de dados apresente melhor estabilidade. Esta métrica apresentou que o melhor resultado para os conjuntos são 2 agrupamentos (Figura 4.15 e 4.16)

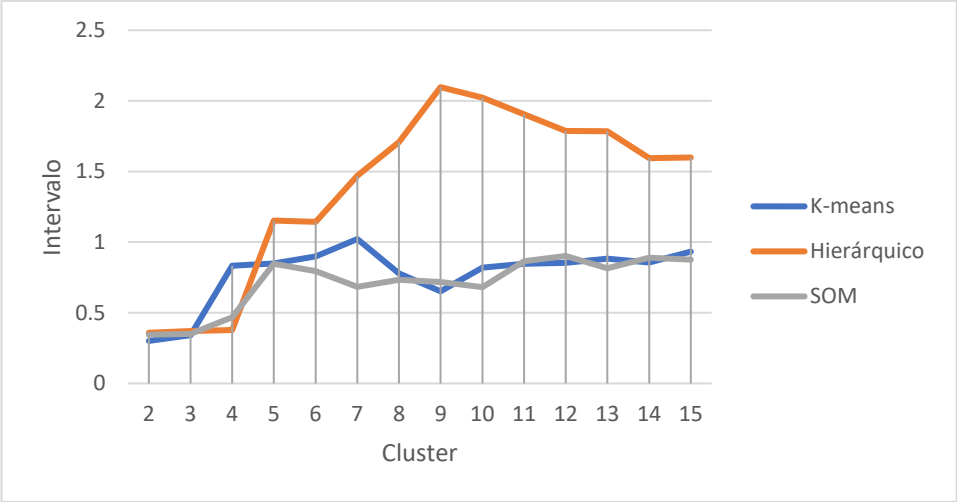


Figura 4.15: Métrica AMD para o conjunto de dados da linhagem CP258, que mostra que a melhor quantidade de agrupamento é 2, onde os dados mostram melhor estabilidade

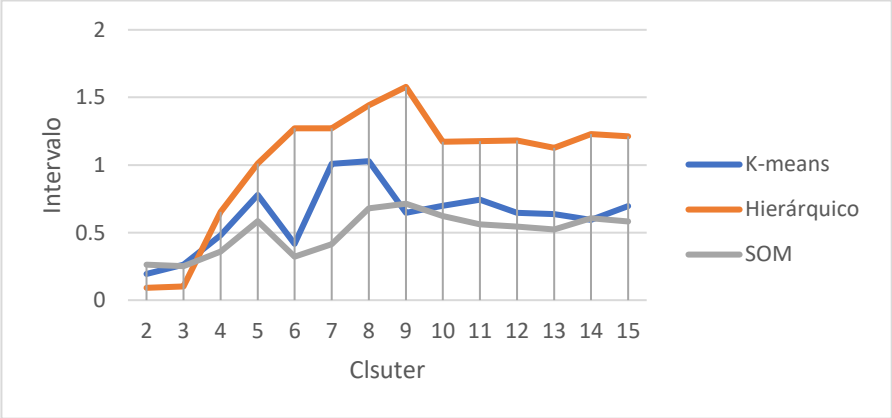


Figura 4.16: Métrica AMD para conjunto de dados da linhagem CP1002, que mostra que a melhor quantidade de agrupamentos é 2, onde os dados mostram melhor estabilidade.

4.3 Clustering

4.3.1 Utilização dos algoritmos de agrupamentos

A partir dos resultados das métricas foram implementados os algoritmos de agrupamentos selecionados pelo modelo com os dados de expressão analisados. Foi implementado o algoritmo K-means com número de cluster (K) igual a K=2, baseado na distância Euclidiana (equação 10). O algoritmo hierárquico foi implementado com a metodologia aglomerativa (down-up), com a técnica means linkage para 2 agrupamentos. SOM foi implementado com uma altura igual 1 e com neurônio com pesos igual a 2, gerando dois agrupamentos para este algoritmo.

Para a obtenção e análise dos agrupamentos, foram usados softwares de mineração de

dados Weka (HALL et al., 2009) e a linguagem R (com pacotes especializados Cluster, Clust, factorextra e SOM, para a obtenção e análises de agrupamentos (IHAKA, ROSS ; GENTLEMAN, 1996).

A Tabela 4.2 mostra a tamanho dos agrupamentos formados pelos diferentes algoritmos. As figuras seguintes mostram a distribuição gráfica dos agrupamentos K-means e hierárquico e SOM para as duas linhagens.

Tamanhos dos agrupamentos Cp258			
Algoritmos	**Hierárquico**	**K-Means**	**Kohonen**
Cluster 1	1025	664	678
Cluster 2	149	510	496
Tamanhos dos agrupamentos Cp258			
Algoritmos	**Hierárquico**	**K-Means**	**Kohonen**
Cluster 1	969	653	704
Cluster 2	160	475	425

Tabela 4.2: Tamanho dos agrupamentos de dados segundo o tipo de algoritmo usados, nas diferentes linhagens

Para a avalição de dos agrupamentos formados pelos algoritmos foi utilizado o índice Jaccard o que possibilitou comparar a similaridade das instancias contidas em cada agrupamento, sendo que os agrupamentos formados no CP 258 apresentaram uma similidade entre 52-95%, já CP1002 apresentou similaridade entre 57-84%.

4.3.2 Resultados dos algoritmos de agrupamentos

4.3.2.1 Resultados do algoritmo k-means

Nesta seção, são apresentados os resultados obtidos pelo algoritmo K-means para as linhagens CP258 e CP1002 (Figuras 4:17- 4.20). Sendo que as fronteiras de agrupamentos foram bem definidas por este algoritmo para os conjuntos utilizados.

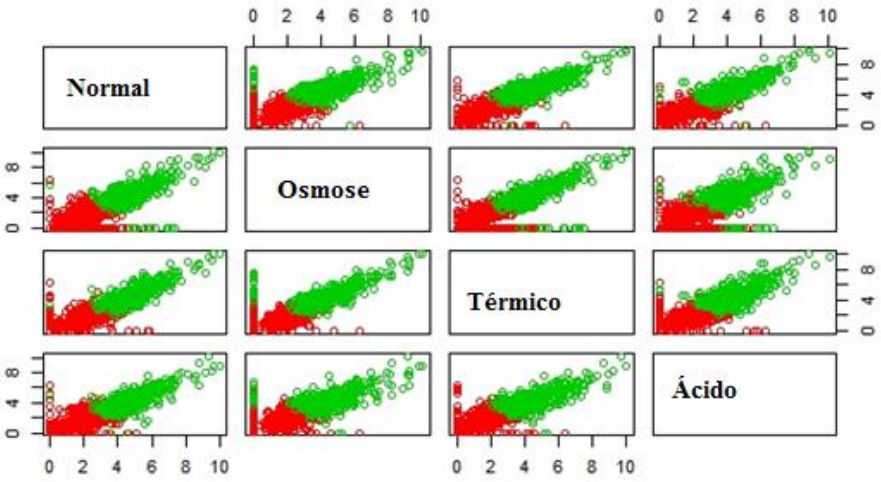


Figura 4.17: Resultado do algoritmo K-means, para a linhagem 258, para todas as condições de estresse a que foi submetido o microrganismo

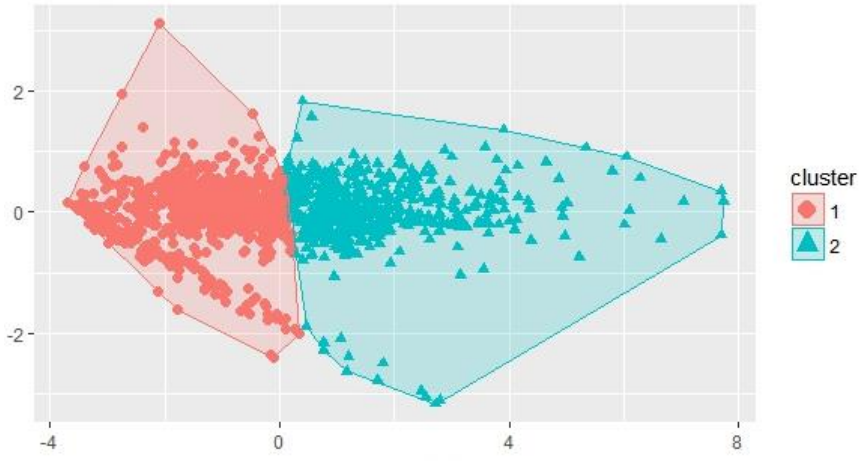


Figura 4.18: Visão geral dos dois agrupamentos formados pela linhagem CP258 com o algoritmo K-means.

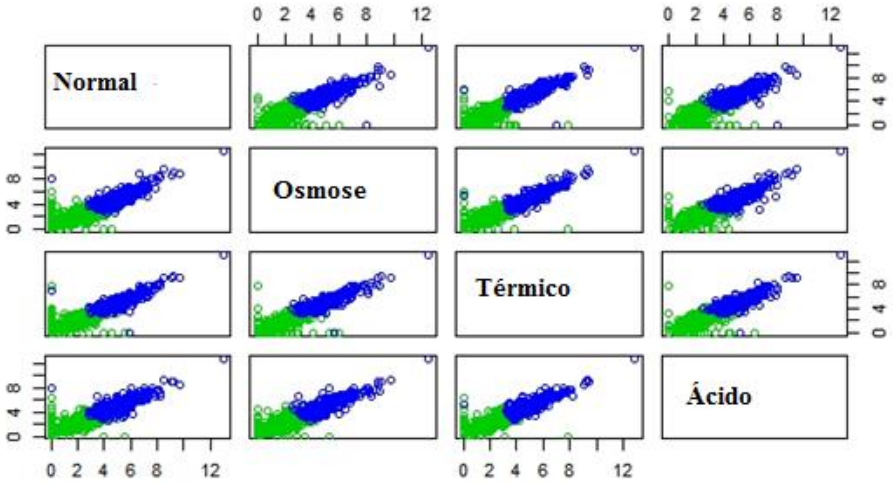


Figura 4.19: Resultado do algoritmo K-means, para a linhagem 1002, para todas as condições de estresse a que foi submetido o microrganismo

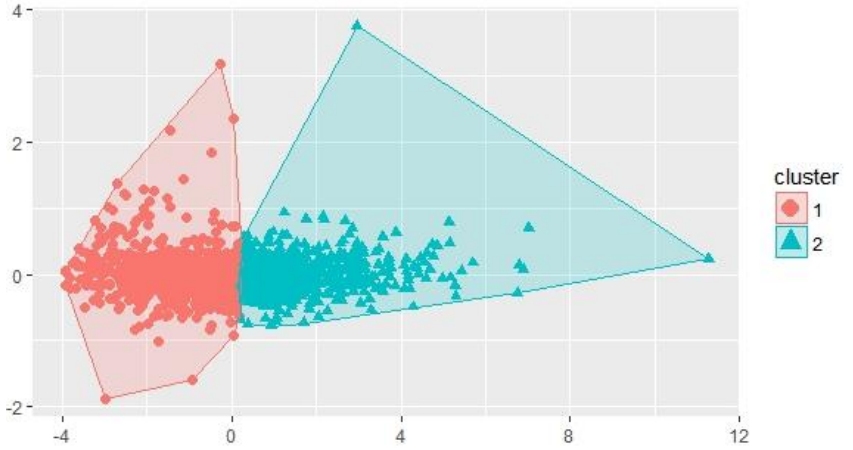


Figura 4.20: Visão geral dos dois agrupamentos formados pela linhagem CP258 com o algoritmo K-means.

4.3.2.2 Resultados do algoritmo hierárquico

Nesta seção, são apresentados os resultados obtidos pelo algoritmo hierárquico para as linhagens CP258 e CP1002 (Figuras 4:21- 4.24). Este algoritmo mostrou uma superposição entre os dois agrupamentos que foram encontrados.

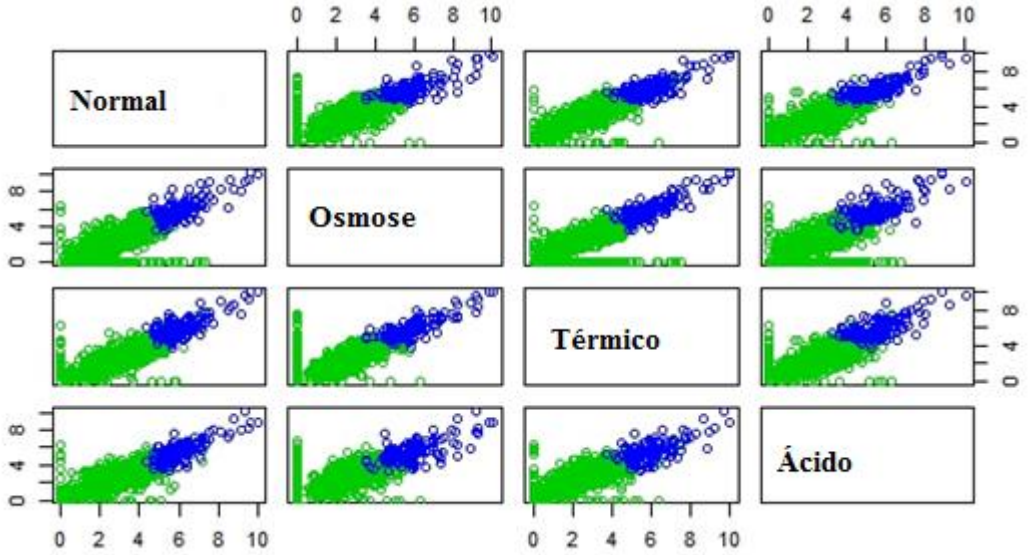


Figura 4.21: Resultado do algoritmo hierárquico, para a linhagem CP258 para todas as condições de estresses que foi submetida o microrganismo

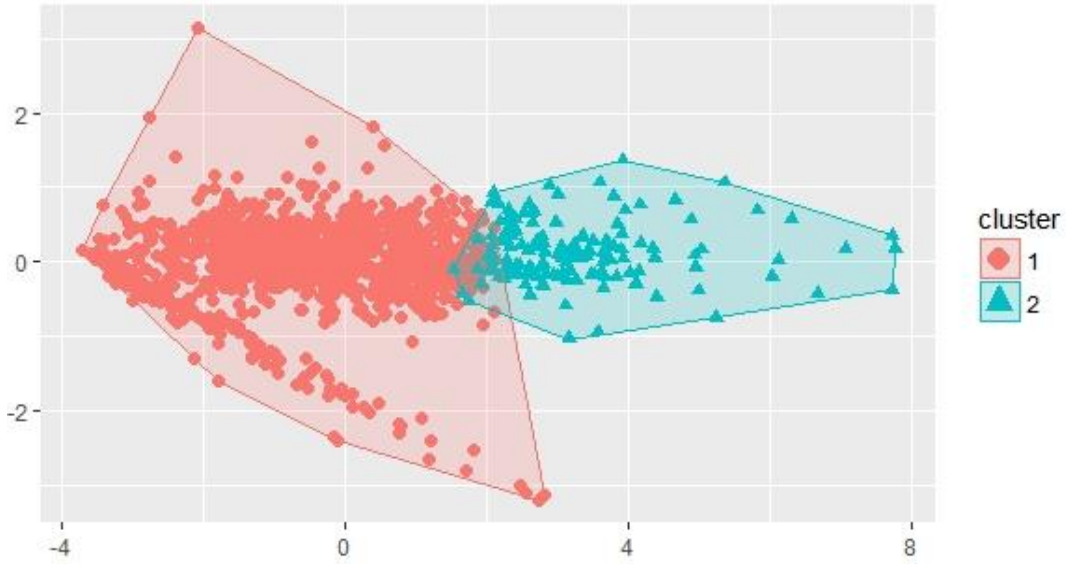


Figura 4.22: Visão geral dos dois agrupamentos formados pela linhagem CP258 com o algoritmo hierárquico.

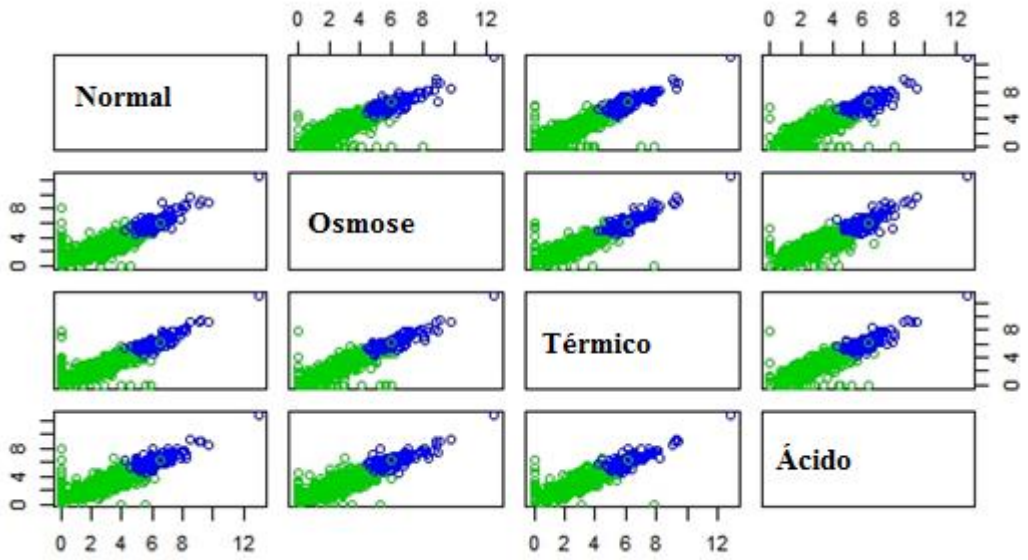


Figura 4.23: Resultado do algoritmo hierárquico, para a linhagem CP1002 para todas as condições de estresses que foi submetida o microrganismo

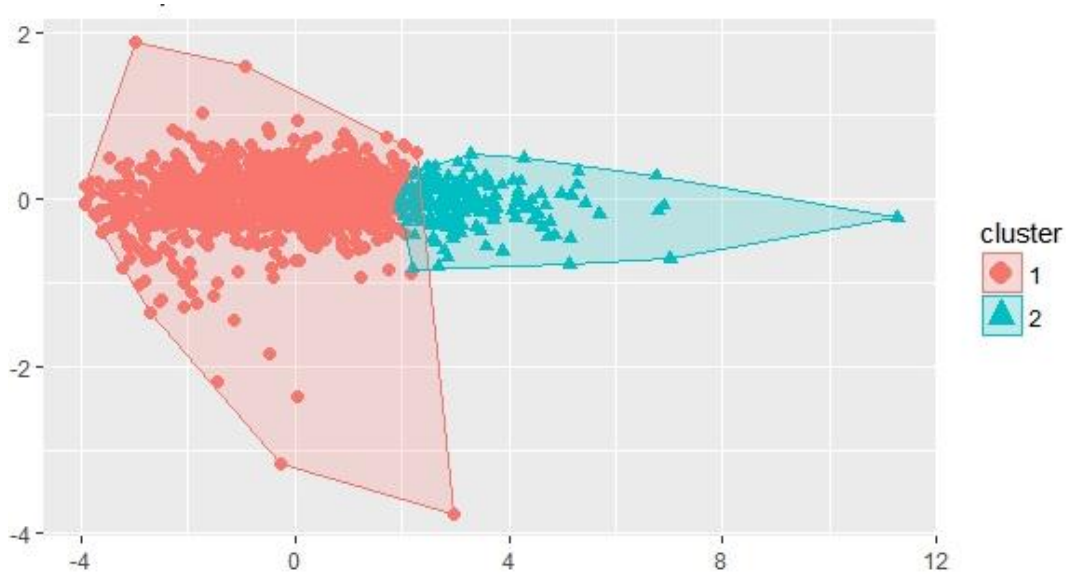


Figura 4.24: Visão geral dos dois agrupamentos formados pela linhagem CP1002 com o algoritmo hierárquico.

4.3.2.3 Resultados do algoritmo SOM

Nesta seção, são apresentados os resultados obtidos pelo algoritmo *Self Organization Map* (SOM) para as linhagens CP258 e CP1002 (Figuras 4:21- 4.24). Este algoritmos foi capaz de definir boas fronteiras de separação entres os agrupamentos formados.

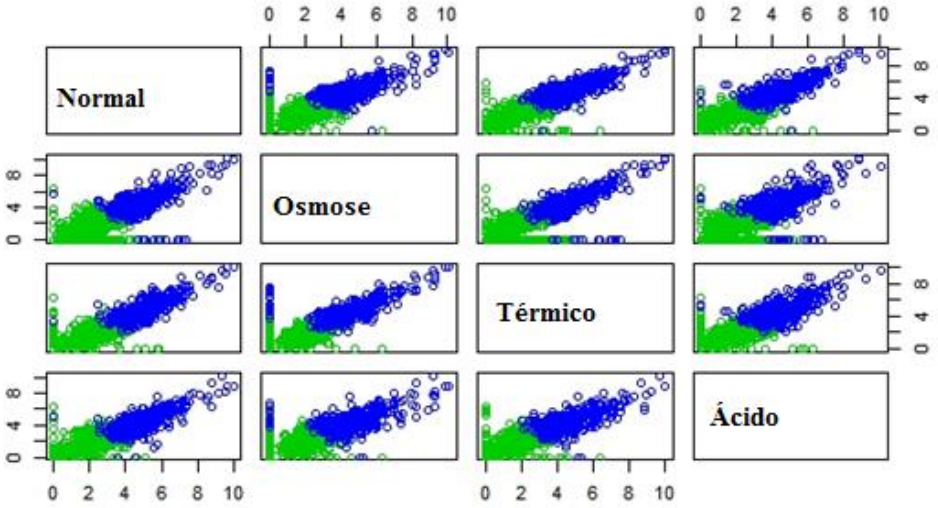


Figura 4.25: Resultado do algoritmo SOM, para a linhagem CP258 para todas as condições de estresses que foi submetida o microrganismo

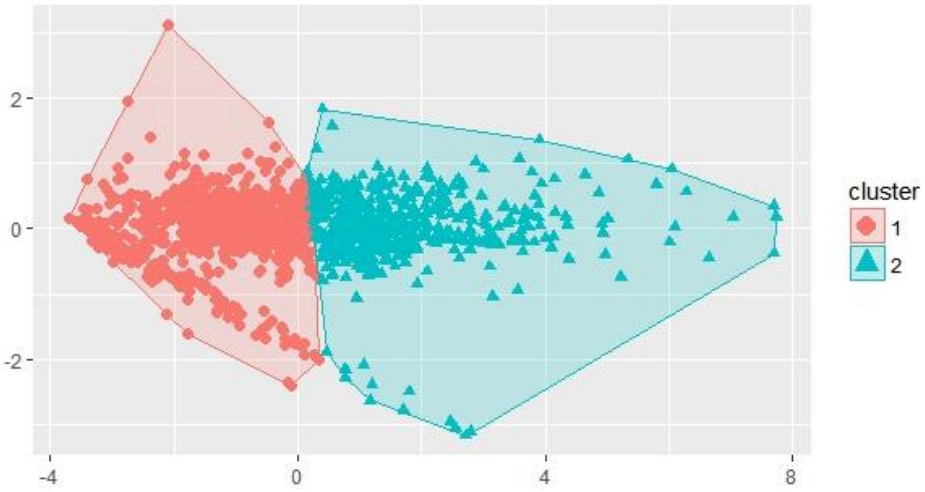


Figura 4.26: Visão geral dos dois agrupamentos formados pela linhagem CP2501002 com o algoritmo SOM

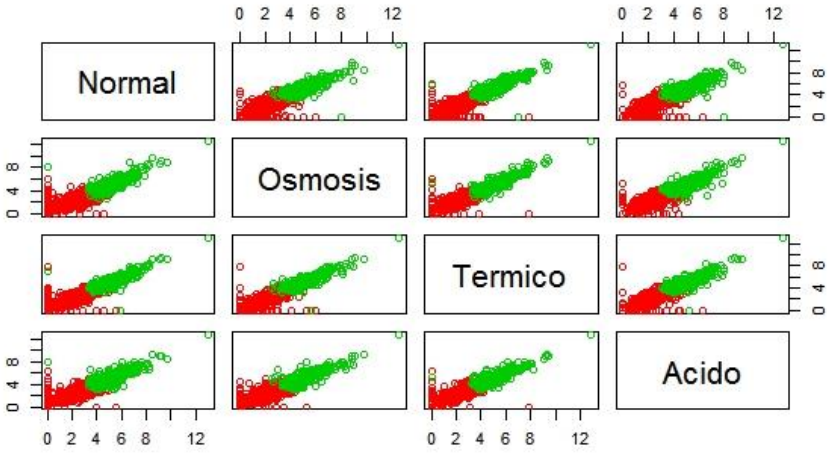


Figura 4.27Resultado do algoritmo SOM, para a linhagem CP1002 para todas as condições de estresses que foi submetida o microrganismo.

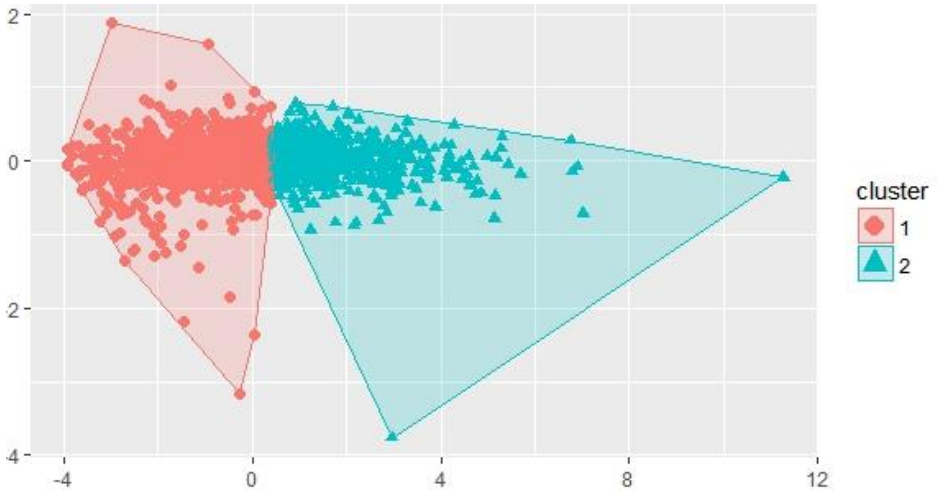


Figura 4.28: Visão geral dos dois agrupamentos formados pela linhagem CP2501002 com o algoritmo SOM

4.3.3 Similaridade entres os agrupamentos

Para verificar a similaridade entre os agrupamentos, foi usado o método heurístico descrito na seção 3.3.3, o que possibilitou a comparação entre os agrupamentos similares que foram obtidos pelos diferentes algoritmos. Isto permite que os agrupamentos com o mesmo perfil genômico sejam comparados aos de perfil igual ou similar. A tabela 4.3 mostra os perfis da média dos agrupamentos que permitem fazer a comparação de similaridade com base nos perfis de agrupamentos.

Perfil dos agrupamentos do genoma 258			
Agrupamentos	**Hierárquico**	**K-Means**	**Kohonen**
Cluster 1	2.7	2.5	2.08
Cluster 2	5.8	4.5	4.5
Perfil dos agrupamentos do genoma 1002			
Agrupamentos	**Hierárquico**	**K-Means**	**Kohonen**
Cluster 1	2.9	2.3	2.4
Cluster 2	6.13	4.9	5.04

Tabela 4.3 Para cada um dos agrupamentos gerados pelos algoritmos foi calculador o perfil baseado na média aritmética de cada um dos agrupamentos para a comparação dos grupos e sua similitude.

4.4 Pós-Clustering

4.4.1 Lista de genes housekeeping a usada como referência.

Para este microrganismo *Corynebacterium pseudotuberculosis*, foram selecionados 17 genes housekeeping (quadro 4.1) identificados na literatura como genes candidatos. A seleção foi baseada no estudo de Rocha et al. (ROCHA; SANTOS; PACHECO, 2015a) que

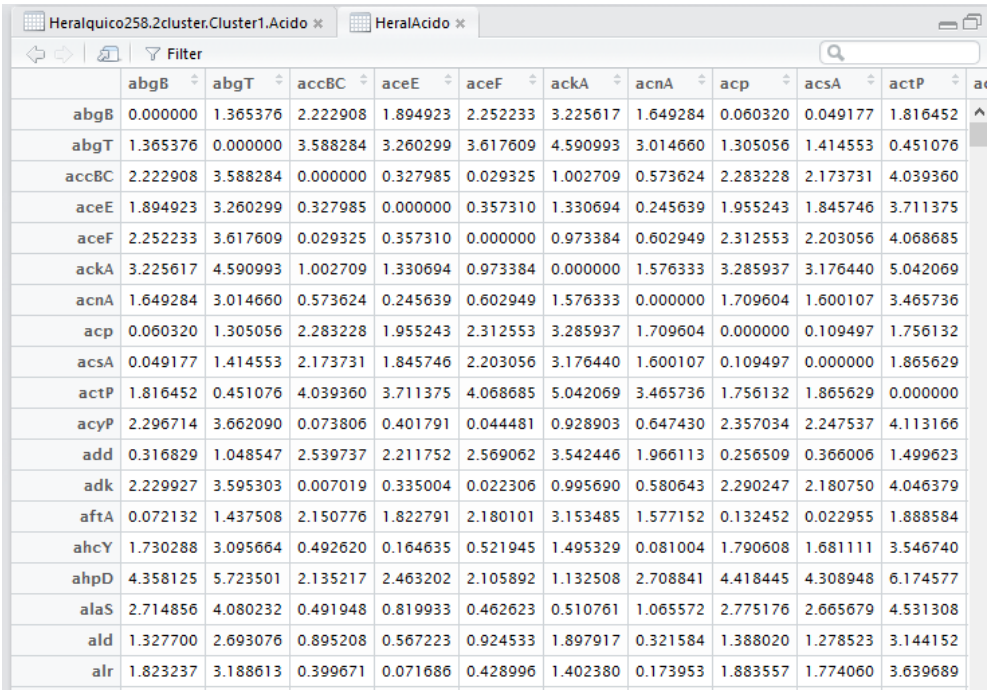
identificaram um conjunto de genes housekeeping de referência para estudos em bactérias através de uma revisão na literatura. Os genes selecionados foram validados por estudos de validação de RT-qPCR, como genes de referência ou housekeeping para bactérias. Para seleção desta lista final, os genes tinham que ser validados por dois ou mais estudos com a técnica, além de apresentarem uma estabilidade em todas as condições de teste.

Genes Housekeeping
adk
dnaG
ftsZ
gap
gmk
gyrB
recA
secA
rho
rpoA
rpoB
rpoC
recF
fusA
glnA
tuf
gyrA

Quadro 4.1 Lista de HKG validados por estudos de RT-qPCR, selecionado para o estudo com *Corynebacterium pseudotuberculosis*

4.4.2 Obtenção das matrizes de distâncias

Para a obtenção dos genes que serão considerados como candidatos, foram utilizadas matrizes estabelecidas por meio das distâncias euclidianas (equação 10), a partir do método descrito na seção 3.4.2. Baseado neste método foram obtidas 4 matrizes por agrupamento (cluster), para um total de 8 matrizes por algoritmos usados, e um total geral de 48 matrizes, para os dois conjuntos de dados que estão sendo usados nesta pesquisa. A Figura 4.29 mostra um exemplo de uma das matrizes de distância obtida para uma das condições de estresse de um agrupamento formado por um algoritmo específico.



Heralquico258.2cluster.Cluster1.Acido | HeralAcido

Filter

	abgB	abgT	accBC	aceE	aceF	ackA	acnA	acp	acsA	actP
abgB	0.000000	1.365376	2.222908	1.894923	2.252233	3.225617	1.649284	0.060320	0.049177	1.816452
abgT	1.365376	0.000000	3.588284	3.260299	3.617609	4.590993	3.014660	1.305056	1.414553	0.451076
accBC	2.222908	3.588284	0.000000	0.327985	0.029325	1.002709	0.573624	2.283228	2.173731	4.039360
aceE	1.894923	3.260299	0.327985	0.000000	0.357310	1.330694	0.245639	1.955243	1.845746	3.711375
aceF	2.252233	3.617609	0.029325	0.357310	0.000000	0.973384	0.602949	2.312553	2.203056	4.068685
ackA	3.225617	4.590993	1.002709	1.330694	0.973384	0.000000	1.576333	3.285937	3.176440	5.042069
acnA	1.649284	3.014660	0.573624	0.245639	0.602949	1.576333	0.000000	1.709604	1.600107	3.465736
acp	0.060320	1.305056	2.283228	1.955243	2.312553	3.285937	1.709604	0.000000	0.109497	1.756132
acsA	0.049177	1.414553	2.173731	1.845746	2.203056	3.176440	1.600107	0.109497	0.000000	1.865629
actP	1.816452	0.451076	4.039360	3.711375	4.068685	5.042069	3.465736	1.756132	1.865629	0.000000
acyP	2.296714	3.662090	0.073806	0.401791	0.044481	0.928903	0.647430	2.357034	2.247537	4.113166
add	0.316829	1.048547	2.539737	2.211752	2.569062	3.542446	1.966113	0.256509	0.366006	1.499623
adk	2.229927	3.595303	0.007019	0.335004	0.022306	0.995690	0.580643	2.290247	2.180750	4.046379
aftA	0.072132	1.437508	2.150776	1.822791	2.180101	3.153485	1.577152	0.132452	0.022955	1.888584
ahcY	1.730288	3.095664	0.492620	0.164635	0.521945	1.495329	0.081004	1.790608	1.681111	3.546740
ahpD	4.358125	5.723501	2.135217	2.463202	2.105892	1.132508	2.708841	4.418445	4.308948	6.174577
alaS	2.714856	4.080232	0.491948	0.819933	0.462623	0.510761	1.065572	2.775176	2.665679	4.531308
ald	1.327700	2.693076	0.895208	0.567223	0.924533	1.897917	0.321584	1.388020	1.278523	3.144152
alr	1.823237	3.188613	0.399671	0.071686	0.428996	1.402380	0.173953	1.883557	1.774060	3.639689

Figura 4.29: Exemplo de matriz de distância, obtida para o cluster 1 do algoritmo hierárquico na condição de estresse ácido. Este é uma matriz 2X2, que indica a distância que temos desde cada um dos genes para outro nesta condição.

Após a obtenção das matrizes de distância, são obtidos os valores das métricas de estatística descritiva das matrizes de distâncias e, a partir destas, são construídas a submatrizes de distâncias. Nesta pesquisa foi selecionado como limiar de corte o primeiro quartil, por apresentar um limiar que permite identificar os genes que estão mais próximos dos genes housekeeping de referência. A figura mostra um exemplo das métricas estatísticas, obtidas a partir das matrizes de distância, com a função de *summary* da suíte R.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.5987  1.2650  1.4780  2.1450  6.7030
```

Figura 4.30: Exemplo das métricas de estatísticas descritivas das matrizes de distancias. Este resultado o primeiro quartil que está é usado como limiar para a construção das submatrizes de distancias.

4.4.3 Criação das submatrizes baseadas no limiar de corte.

Para a redução do hiperplano de dados, foram construídas submatrizes a partir das matrizes de distâncias, usando como limiar de corte o primeiro quartil. Isto possibilitou a redução das matrizes entre 70% e 90% dos genes originais, conservando para a pesquisa só os

genes mais próximos aos genes de referência. Neste passo foram construídas 48 submatrizes, com tamanho de 10% a 30% do valor original dos conjuntos de dados.

4.4.4 Análise dos possíveis candidatos por agrupamentos e conjunto de dados

Após a obtenção das submatrizes de distâncias, foram construídas as matrizes, através da combinação de genes que apresentaram proximidade com relação a cada uns dos genes de referência nas diferentes condições, para os dois conjuntos de dados. Estas matrizes possibilitaram obter os genes que apresentaram 3 e 4 coincidências de proximidade nas diferentes condições de estresse. A Tabela apresenta a quantidade de genes que foram identificados nas diferentes condições, segundo os genes usados como referência.

Após obter os genes com 3 e 4 coincidências de condições nas condições de estresse, foram identificados os genes constantes nos diferentes algoritmos. Os genes que foram constantes nos diferentes algoritmos e conjuntos de dados, com proximidade a um HKG validado, foram selecionados como os candidatos a genes housekeeping. Neste passo, foram obtidos 32 genes como possíveis candidatos a housekeeping.

Genes de Referências	Hierárquico			K-means			SOM		
	4 coincidências	3 coincidências	3 e 4 coincidências	4 coincidências	3 coincidências	3 e 4 coincidências	4 coincidências	3 coincidências	3 e 4 coincidências
Dnag	5	14	19	1	4	6	1	5	6
Glna	2	8	5	0	3	2	0	4	2
Gyra	0	10	8	0	2	0	0	2	0
Recf	1	19	18	0	6	5	0	8	5
Adk	0	0	0	0	1	0	0	0	0
Fusa	1	1	0	1	1	0	1	1	0
Gap	1	2	0	1	3	0	1	3	0
Gmk	0	0	0	0	0	0	0	0	0
Gyrb	5	9	0	1	4	0	1	4	0
Reca	3	8	20	0	3	2	0	3	5
Rho	8	5	0	0	0	0	0	3	0
RpoA	0	3	1	0	7	2	0	7	2
RpoB	2	6	7	2	7	0	1	7	7
RpoC	0	0	0	1	4	0	2	2	0
Seca	13	4	5	0	2	7	0	2	7
Tuf	1	0	0	1	0	1	1	0	1
Ftsz	0	0	0	0	0	0	0	0	0

Tabela 4.4: Quantidade de genes selecionados que apresentaram coincidências nas diferentes condições de estresse, com relação aos algoritmos utilizados e os genes usados como referências.

Os genes obtidos foram classificados em três grupos, 1) os genes que apresentaram

proximidade com um HKG validado nas 4 condições de estresse estudadas nas duas linhagens usadas 2) genes que mantiveram proximidade com relação ao gene validado de referência em 3 das condições de estresse, nas duas linhagens; 4) os genes que apresentaram proximidade com relação ao HKG de referência em 4 ou 3 condições em uma das duas linhagens

Genes com 4 coincidência	**Genes com 3 coincidência**	**Genes com 3 e 4 coincidências**
dnaA	aftA	hemB
ispG	lysI	pafA1
gpmA	ctpA	pcrA
	pbpB	carA
	pcsA	dapD
	pepC2	glmU
	tcsS3	helZ
	tcsS5	mfd
	ndk	pgk
	ppc	sigH
	wbbL	ligA
	deoR	ackA
	rplC	fxsA
	rplD	efp
	hemA	
	lysS	
	merR2	

Quadro 4.2: Genes que apresentaram coincidências nos diferentes algoritmos usados, que foram selecionados como possíveis genes housekeeping

4.4.5 Avalição da estabilidade dos possíveis genes candidatos a HKG

Após a identificação, os genes foram submetidos a um processo de filtragem a partir de técnicas estatísticas (coeficiente de variação (equação 12), *máximo change fold* (equação 13) e diferença de coeficiente de variação (equação 14), onde foram selecionados os genes com melhor estabilidade dentro dos limiares estabelecidos como candidatos finais a genes housekeeping.

Para os coeficientes de variação nas duas linhagens, foi estabelecido como limiar de corte os valores de variação menores a 10% entres os genes nas duas linhagens, obtendo como resultados os genes candidatos com expressão mais estável. Os genes foram avaliados por meio do MFC e foram selecionados os que mostraram menor variação de expressão entres as diferentes condições. A diferença Dcv permitiu avaliar a variação entre as diferentes linhagens, sendo selecionados os genes que apresentaram uma diferença menor que 5% entre a expressão das duas linhagens.

Esta avalição de estabilidade permitiu a identificação de 17 genes (Tabela 5) que são propostos como possíveis candidatos a housekeeping.

	258				1002				
Genes com 4 coincidências									
Genes	**σ**	**$\overline{X}$**	**CV**	**MFC**	**σ**	**$\overline{X}$**	**CV**	**MFC**	***Dcv***
ispG	0.21	4.15	5.1%	1.13	0.19	4.84	4.02%	1.10	1.1%
gpmA	0.43	5.74	7.5%	1.22	0.27	6.10	4.39%	1.13	3.1%
Genes com 3 coincidências									
Genes	**σ**	**$\overline{X}$**	**CV**	**MFC**	**σ**	**$\overline{X}$**	**CV**	**MFC**	***Dcv***
tcsS3	0.23	2.62	8.7%	1.27	0.26	2.77	9.54%	1.26	0.8%
ndk	0.52	5.14	10.0%	1.30	0.36	5.34	6.70%	1.17	3.4%
wbbL	0.27	4.03	6.7%	1.17	0.27	4.24	6.35%	1.18	0.3%
ppc	0.3	3.5	7.90%	1.3	0.1	4.3	3.23%	1.1	4.7%
deoR	0.32	3.90	8.2%	1.20	0.16	4.08	4.03%	1.11	4.2%
rplB	0.24	4.95	4.8%	1.12	0.18	5.49	3.33%	1.09	1.5%
rplD	0.29	5.35	5.4%	1.15	0.29	5.61	5.21%	1.13	0.2%
lysS	0.13	4.39	3.0%	1.08	0.16	4.65	3.54%	1.09	0.5%
Genes com 3 e 4 coincidência									
Genes	**σ**	**$\overline{X}$**	**CV**	**MFC**	**σ**	**$\overline{X}$**	**CV**	**MFC**	***Dcv***
helZ	0.24	2.22	10.6%	1.33	0.19	2.89	6.63%	1.19	4.0%
mfd	0.18	2.62	6.9%	1.19	0.17	3.05	5.60%	1.15	1.3%
sigH	0.25	3.74	6.6%	1.18	0.38	4.68	8.21%	1.26	1.6%
ligA	0.23	3.92	5.9%	1.16	0.17	4.10	4.09%	1.12	1.8%
fxsA	0.29	4.49	6.5%	1.18	0.36	4.66	7.76%	1.18	1.2%
efp	0.22	4.23	5.1%	1.13	0.22	4.53	4.84%	1.13	0.3%

Tabela 4.5: Avalição da estabilidade dos genes candidatos a HKG, por meio do desvio padrão, média, coeficiente de variação e diferença de CV entre linhagens dos genes selecionados como candidatos a housekeeping

Os genes selecionados como candidatos a HKG mostram expressão estável nas diferentes condições a que foram submetidos, mostrando pouca variação, o que pode ser uma evidência de que estes genes podem ser confirmados como candidatos a HKG através dos experimentos de bancada. A figura mostra os heatmap com a expressão dos genes selecionados como candidatos a housekeeping, nas duas linhagens que foram utilizadas.

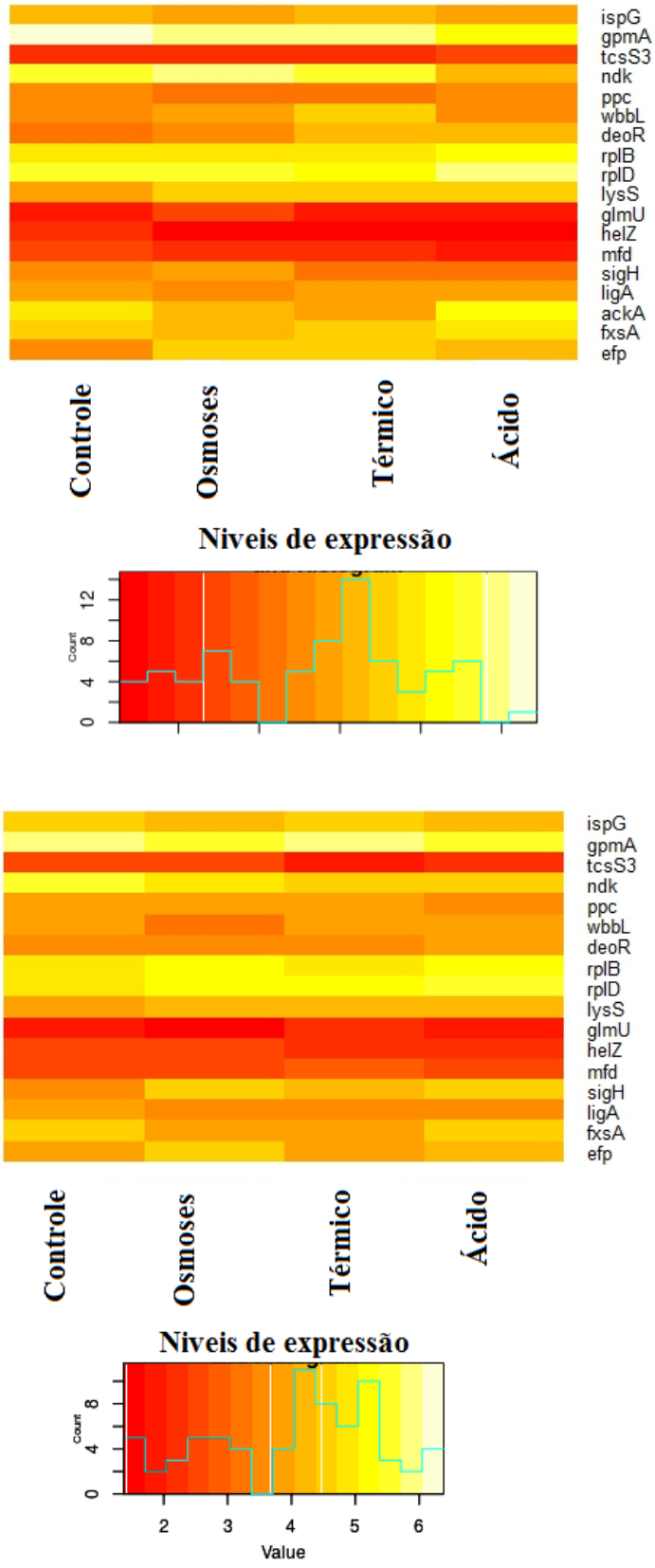


Figura 4.31: Encima CP258 e embaixo Cp1002, a imagem mostra os perfis de expressão dos genes selecionados como candidatos nas duas linhagens utilizadas, sendo que estes genes apresentam estabilidade constante nas diferentes linhagens

Os genes selecionados apresentam diferentes funções biológicas dentro dos genomas. A fim de identificar essas funções, foram feitas análises dos processos biológicos com a plataforma Gene Ontology (GO). Esta analises permite identificar os processos referente ao objetivo biológico para qual o gene ou produto genético contribuiu e podem ser realizados por uma o mais vias (THE GENE ONTOLOGY CONSORTIUM, 2016). O conjunto de genes foi classificado em 4 funções biológicas (figura 4.32), tomando como microrganismo de referência a *Mycobaterium tuberculoses*, já que é o organismo disponível no banco de dados filogeneticamente mais próximo. Dos genes estudados, 50%, foram classificados com função biológica de atividade catalítica; 29% foram classificados no processo ligação genômica dentro do microrganismo; 14% dos genes participam em atividades relacionadas com a estrutura molecular do organismo e 7% participaram em vias relacionadas com os processos ligados à regulação da tradução.

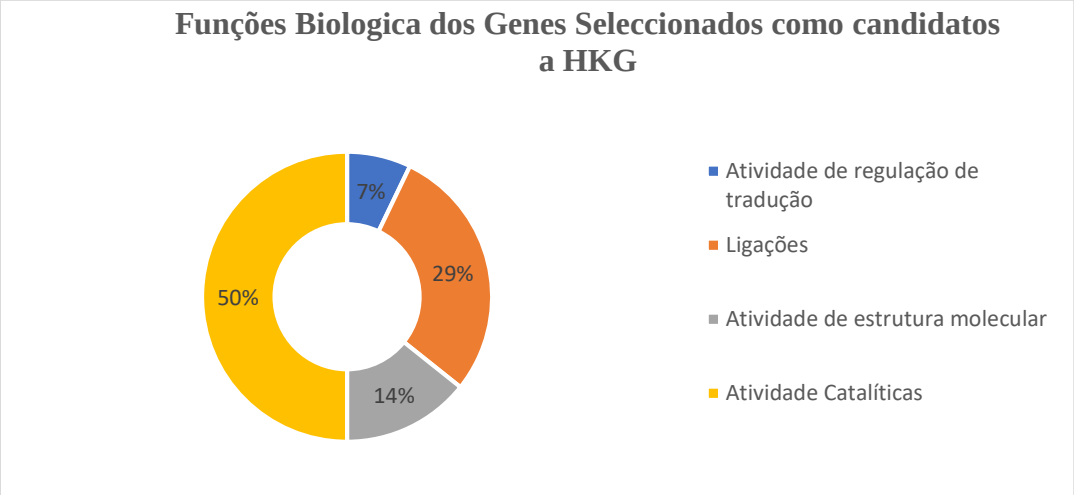


Figura 4.32: Funções biológica dos genes selecionados como HKG pela abordagem, baseado no analises de GO, usando como referência *Mycobaterium Tuberculoses como* organismo de referência.

Da mesma forma, foram identificadas a funções específicas de cada um dos genes identificados pela abordagem como candidatos a genes housekeeping. O quadro descreve as funções que estes genes desempenham dentro do genoma da Corynebacterium pseudotuberculosis. Isto permite uma compreensão mais clara dos genes.

Genes	Descrição das Funções Biológicas dos Candidatos a HKG
ispG	4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase
gpmA	Phosphoglycerate mutase activity
tcsS3	Two-component system sensor kinase
ndk	Nucleoside diphosphate kinase
wbbL	dTDP-Rha:alpha-D-GlcNAc-pyrophosphate polyprenol, alpha-3-L-rhamnosyltransferase
ppc	Phosphoenolpyruvate carboxylase
deoR	Deoxyribonucleoside regulator
rplB	50S ribosomal protein L2
rplD	50S ribosomal protein L4
lysS	lysyl-tRNA synthetase
helz	Helicase helZ
mfd	Transcription-repair coupling factor
sigH	RNA polymerase sigma-H factor
ligA	NAD-dependent DNA ligase LigA
fxsA	FxsA cytoplasmic membrane protein
efp	Elongation factor P

Quadro 4.3 descrição das funções dos genes selecionado como possíveis candidatos a HKG dentro dos genomas da *Corynebacterium pseudotuberculosis*

5. DISCUSSÃO

O processamento dos dados permitiu a identificação dos candidatos a genes housekeeping, os quais são essenciais, por meio do core genoma. Este conjunto de dados foi submetido a diferentes algoritmos de agrupamentos, o que possibilitou a obtenção dos candidatos finais a HKG para o organismo estudado.

Os algoritmos foram avaliados com diferentes técnicas de avaliação, sendo que os melhores resultados foram obtidos pelo algoritmo hierárquico para a métrica Silhouette com um valor de 0.62 para Cp258 e 0.77 para Cp1002. Este algoritmo também apresentou bom desempenho com relação às demais métricas, como Dunn, DB e conectividade, para os dados nas duas linhagens. Os algoritmos K-means e SOM apresentaram um desempenho similar com relação à avaliação destas métricas. Baseado nos dados, foi feito um consenso dos resultados das diferentes métricas, com intuito de obter a melhor quantidade de agrupamentos, sendo selecionado o valor com maior frequência, que neste caso foi de 2 agrupamentos para cada linhagem.

Neste estudo, foi evidente que a utilização de métricas de avalição é de vital importância para o entendimento dos dados, esta permitem poder determinar as tendências dos dados, assim como identificar a priori as possíveis quantidades de agrupamentos, que após devem ser verificados com os algoritmos. Foram utilizadas várias técnicas para poder obter um consenso por votação da melhor quantidade de agrupamentos presentem nos conjuntos de dados utilizados.

Para um melhor funcionamento dos algoritmos, transformando os dados para log +1, o que possibilitou obter um conjunto de dados mais próximo da normalidade. Com isso, observamos que algoritmos de agrupamentos baseados na distância mostraram bom desempenho na definição da fronteira de separação dos agrupamentos, sendo que os melhores resultados foram obtidos para os dados de CP258 por SOM e para CP 1002 K-means e SOM. Cabe ressaltar que o conjunto de dados utilizados apresentaram atributos com superposição, o que dificultou a definição das fronteiras de separação. No caso da CP258, por exemplo, houve problemas de superposição nos dados de estresse osmótico e condição normal. Já no CP1002, os dados de estresse térmico e osmótico também apresentou dificuldades de superposição de instancias o que dificultou definir as fronteiras de separação dos agrupamentos.

.

É importante reiterar que nesta pesquisa a equipe não utilizou o *ensembler clustering*. Pretendemos utilizar esta técnica de clustering em trabalhos futuros, com o objetivo de comparar seu desempenho com relação à abordagem proposta nesta pesquisa, a fim de identificar possíveis candidatos a HKG.

A utilização de duas linhagens de *Corynebacterium pseudotuberculosis* possibilitou a comparação dos genes obtidos com candidatos, por meio das técnicas de similaridades que foram desenvolvidas para comparar os perfis gênicos dos agrupamentos gerados pelos algoritmos. A abordagem possibilitou a obtenção de 34 possíveis candidatos a genes housekeeping (quadro 4.2). Para a seleção da lista final, foi realizada uma filtragem baseada em técnica estatística apresentada em De Jonge et al.,(2007), que apresenta índices de coeficientes de variação e máximo change fold como métricas para a seleção dos candidatos a genes housekeeping.

Tendo essa informação como base, para este trabalho determinamos como limiar a cv <10 e, para obter este limiar, foi feita uma relação gráfica entre os CV das diferentes linhagens, sendo selecionado 10 como limiar de corte, já que é onde as duas métricas se curtam. Já o MCF <2 foi selecionado a partir da pesquisa de Carvalho et al. (2014b), na qual os genes com um valor maior a 2 são considerados superexpressos e com pouca estabilidade. Além disso, foi feita uma comparação entre os resultados das duas linhagens a partir de Dcv e foram considerados, através de diferentes testes, que o melhor limiar de variação entre Cv das linhagens foi menor que 5%, sendo que este permitiu obter os genes com maior estabilidade dos diferentes conjuntos de dados. Através destes limiares das métricas, pode-se obter uma lista final com 17 genes que podem ser considerados candidatos a HKG em *Corynebacterium pseudotuberculosis*.

Um fato importante é que 6 dos genes identificados como housekeeping neste organismo já foram descritos por pelo menos uma pesquisa como possível candidato os genes HKG ou genes de grande importância em outros organismos. O gene Gpma é como HKG conservado no *Lactobacillus acidophilus* (RAMACHANDRAN et al., 2013); o gene IspG é essencial para o crescimento in vitro da *Mycobacterium tuberculosis* (BROWN; KOKOCZKA; PARISH, 2015). Este gene também é associado com sobrevivência intracelular e a indução de respostas imunitárias em outro patógenos bacterianos (GAHAN; HILL, 2012), assim como há

hipóteses de que o gene pode ser essencial para a patogênese de *Listeria monocytogenes* (ABDELHAMED; LAWRENCE; KARSI, 2015); o gene rplD foi validado como um gene de referência para *Staphylococcu*s aureus sub diferentes condições de estresse (SIHTO et al., 2014) e avaliado como HKG em *flavobacterium Zobellia galactanivorans* (THOMAS; BARBEYRON; MICHEL, 2011); o gene lysS em *E.Coli* ativa isoacceptor de tRNA, este gene codifica a sínteses housekeeping, assim como a interrupção deste pode causa um fenótipo sensível ao frio, é expresso constitutivamente e está sujeito ao controle da taxa de crescimento (PUTZER; LAALAMI, 2000); o gene ppc foi identificado por duas pesquisa como candidato a housekeeping na *cyanobacteria* (PINTO et al., 2012) (SZEKERES et al., 2014); o gene efp foi identificado como candidato a HKG na *Corynebacterium pseudotuberculosis* usando dados de RNA-seq (CARVALHO et al., 2014b). Estes genes foram identificados na literatura como candidatos a HKG e foram avaliados em artigos publicados anteriormente. Devemos ressaltar que a identificação destes genes em diferentes pesquisas como candidatos HKG demostra a eficiência da abordagem que está sendo proposta, para a identificação de HKG em dados de RNA-seq.

Dois genes que não foram definidos na literatura como HKG, mas podem desempenhar funções importantes dentro do genoma da *Corynebacterium pseudotuberculosis* são os genes: helz e sigH. O primeiro é uma helicase é responsável por catalisar a reação NTP + H_2O = NDP + fosfato, que impulsa o desenrolamento da hélice de DNA e RNA e a nível molecular participa em atividades elementares, como catalise ou ligação (EMBL-EBI, 2017). O gene sigH, codificante do fator sigma H, foi considerado significativo. Estudos realizados em *Mycobaterium tuberculosis* verificou que um mutante para sigH, exibiu susceptibilidade para o estresse térmico e oxidativo in vitro, sugerindo importante papel do sigH na rede regulatória envolvendo reposta ao estresse (RAMAN et al., 2001). E Manganelli et al., (2002) confirmaram em seus estudos que o mutante para sigH foi mais sensível a estresse térmico, porém não mostrou diminuição da capacidade de crescer dentro de macrófagos. Em *Corynebacterium glutamicum*, sigH é responsável pela expressão de Clp, uma protease envolvida na degradação de proteínas não funcionais na célula, normalmente originadas pelo choque térmico (Engels et al., 2004).

Os genes identificados como possíveis candidatos a HKG estão agrupados em 4 diferentes funções biológicas que são essenciais para o desenvolvimento das bactérias, que são: a regulação de atividade de tradução, ligação de DNA, atividade de estrutura molecular e

atividade catalíticas. Na Tabela 6, são apresentadas as funções e descrições dos genes selecionados dentro dos genomas estudados.

Os genes identificados na literatura representam o 35% do total dos genes selecionados pela abordagem como possível candidatos a genes housekeeping nesta bactéria, os demais genes apresentam os requerimentos mínimos de estabilidades para ser HKG. Dentre este conjunto também existe a possibilidade de conter genes que sejam falsos positivos, o que pode ser confirmado através dos estudos de bancadas.

A abordagem proposta é eficiente para a identificação de genes housekeeping em dados de RNA-seq, mas, para que possa ser utilizada em outros trabalhos, há a necessidade de que esteja adaptada ao contexto de estudo, assim como ao microrganismo que está sendo alvo de pesquisa. Os algoritmos de agrupamento, que permitem definir os perfis dos dados, devem ser avaliados e adaptados, assim como os tipos de dados utilizados. Além disso, as condições de estresse usadas devem ser similares para as linhagens estudadas e é recomendável que a linhagens utilizadas sejam do mesmo organismo, o que garantirá melhor eficiência na identificação dos genes alvos.

Um viés da abordagem é com relação à lista de genes HKG validados que foram selecionados para funcionar como referência na identificação dos candidatos, é recomendável que estes genes tenham sido validados por duas ou mais pesquisas, para reduzir o viés de utilização de genes pouco confiáveis. A eficiência da abordagem é proporcional à estabilidade dos genes selecionados como referências.

6. CONCLUSÕES

Este trabalho apresenta uma abordagem eficiente para a identificação de candidatos a genes housekeeping, baseado em algoritmos de agrupamentos e proximidades com genes housekeeping já identificados na literatura, através de análises de dados de expressão gênica.

A identificação de genes housekeeping é fundamental para realização de estudos de expressão e RT-PCR. Esta pesquisa teve, como foco, uma abordagem computacional para a identificação de candidatos a genes housekeeping, através de dados RNA-seq, utilizando algoritmos e técnicas de clustering, o que representa uma inovação na área devido reduzir o tempo e o custo para a identificação de candidatos a estes tipos de genes.

Neste trabalho foram identificados 17 genes candidatos, sendo que 6 deles já foram identificados e definidos como candidatos HKG para outros organismos. Estes genes mostraram boa estabilidade e pouca variações na expressão nas duas linhagens que foram estudadas, o que é um bom indício de que podem ser confirmados como HKG através de estudos de bancada.

A avaliação da estabilidade e variabilidade dos possíveis candidatos a genes housekeeping é importante para se fazer uma filtragem eficiente e poder obter novo candidatos que cumpram com os requisitos mínimos, para serem classificados como candidatos a housekeeping.

Por meio desta abordagem, pode-se evidenciar que é possível encontrar genes housekeeping agrupados e próximos, em procariotos como acontece em eucarioto. Seria interessante verificar as funções que desempenham os genes que compõem estes agrupamentos identificados.

A abordagem proposta nesta pesquisa pode ser aplicada a outros organismos, sendo que é preciso uma boa avalição e adaptação para o bom funcionamento desta. Esta metodologia pode ser um instrumento de identificação eficiente de possíveis de candidatos genes housekeeping, em diferentes organismos alvo de estudo, o que pode constituir uma economia de tempo e recursos.

Como trabalho futuro, espera-se poder validar os genes obtidos através desta abordagem através de experimentos de bancada no laboratório. Assim com a implementação de abordagem usando técnicas de *ensemble cluster* para a identificação dos agrupamentos. Em adição pretende-se automatizar a abordagem através de uma ferramenta de software online, para estar disponibilizada para a comunidade cientifica.

REFERENCIAS

ABDELHAMED, H.; LAWRENCE, M. L.; KARSI, A. A novel suicide plasmid for efficient gene mutation in Listeria monocytogenes. **Plasmid**, v. 81, p. 1–8, 2015.

ANDRITSOS, P. Data clustering techniques. **Toronto, University of Toronto, Dep. of Computer …**, 2002.

ANSORGE, W. J. Next-generation DNA sequencing techniques. **New biotechnology**, v. 25, n. 4, p. 195–203, 2009.

BERZAL, F. M�todos de agrupamiento: clustering. v. 3, 2005.

BOLSHAKOVA, N.; AZUAJE, F. Cluster validation techniques for genome expression data. **Signal Processing**, v. 83, n. 4, p. 825–833, 2003.

BROCK, G.; PIHUR, V.; DATTA, S. clValid: An R package for cluster validation. **J Stat Softw**, v. 25, n. March 2008, p. 1–32, 2008.

BROWN, A. C.; KOKOCZKA, R.; PARISH, T. LytB1 and LytB2 of mycobacterium tuberculosis are not genetically redundant. **PLoS ONE**, v. 10, n. 8, p. 1–12, 2015.

BRUN, M. et al. Model-based evaluation of clustering validation measures. **Pattern Recognition**, v. 40, n. 3, p. 807–824, 2007.

CARVALHO, D. M. et al. Reference genes for RT-qPCR studies in Corynebacterium pseudotuberculosis identified through analysis of RNA-seq data. **Antonie van Leeuwenhoek**, v. 106, n. 4, p. 605–614, 2014a.

CARVALHO, D. M. et al. Reference genes for RT-qPCR studies in Corynebacterium pseudotuberculosis identified through analysis of RNA-seq data. **Antonie van Leeuwenhoek**, v. 106, n. 4, p. 605–14, 2014b.

CONESA, A. et al. A survey of best practices for RNA-seq data analysis. **Genome Biology**, v. 17, n. 1, p. 13, 2016.

DA SILVA JUNIOR, C.; SASSON, S. **Biología**. São Paulo: Editora Saraiva, 2002.

DALTON, L.; BALLARIN, V.; BRUN, M. Clustering algorithms: on learning, validation, performance, and applications to genomics. **Current genomics**, v. 10, n. 6, p. 430–45, 2009.

DATTA, S.; DATTA, S. Comparisons and validation of statistical clustering techniques for microarray gene expression data. **Bioinformatics**, v. 19, n. 4, p. 459–466, 2003.

DE FERRARI, L.; AITKEN, S. Mining housekeeping genes with a Naive Bayes classifier. **BMC genomics**, v. 7, n. 1, p. 277, 2006.

DE JONGE, H. J. M. et al. Evidence based selection of housekeeping genes. **PLoS ONE**, v. 2, n. 9, p. 1–5, 2007.

DHEDA, K. et al. Validation of housekeeping genes for normalizing RNA expression in real-time PCR. **BioTechniques**, v. 37, n. 1, p. 112–119, 2004.

DONG, B. et al. Predicting housekeeping genes based on fourier analysis. **PLoS ONE**, v. 6, n. 6, p. 1–11, 2011.

DORELLA, F. A. et al. Corynebacterium pseudotuberculosis: microbiology, biochemical properties, pathogenesis and molecular studies of virulence. **Veterinary research**, v. 37, n. 2, p. 201–18, 2006.

EISENBERG, E.; LEVANON, E. Y. Human housekeeping genes, revisited. **Trends in Genetics**, v. 29, n. 10, p. 569–574, 2013.

EMBL-EBI, E. M. B. L. **QuickGO**. Disponível em: <https://www.ebi.ac.uk/QuickGO/GProtein?ac=A0A1L6CZP9>.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. Knowledge Discovery and Data Mining: Towards a Unifying Framework. **Proc 2nd Int Conf on Knowledge Discovery and Data Mining Portland OR**, p. 82–88, 1996.

FREE SOFTWARE FOUNDATION, I. **GNU R**. Disponível em: <http://directory.fsf.org/wiki/R#tab=Overview>.

GAHAN, C. G. M.; HILL, C. Isoprenoid biosynthesis in bacterial pathogens. **Microbiology**, v. 158, n. 2012, p. 1389–1401, 2012.

GUOJUN, GAN; CHAOQUN, MA; JIANHONG, W. **Data clustering: theory, algorithms, and applications**. [s.l.] Siam, 2007.

HALKIDI, M.; VAZIRGIANNIS, M. Clustering validity assessment: finding the optimal partitioning of a data set. **Proceedings 2001 IEEE International Conference on Data Mining**, n. FEBRUARY, p. 187–194, 2001.

HALL, M. et al. The WEKA data mining software. **ACM SIGKDD Explorations**, v. 11, n. 1, p. 10–18, 2009.

HAN, JIAWEI; MICHELINE, KAMBER; JIAN, P. **Data Mining: concepts and tecniques**.

3er. ed. Walthan, MA: Morgan Kaufmann, Elsevier, 2011.

HAN, J.; KAMBER, M.; PEI, J. **Data mining: concepts and techniques**. [s.l: s.n.].

HANDL, J.; KNOWLES, J.; KELL, D. B. Computational cluster validation in post-genomic data analysis. **Bioinformatics**, v. 21, n. 15, p. 3201–3212, 2005.

HERNANDEZ SAMPIERI, R.; FERNANDO COLLADO, C.; BAPTISTA LUCIO, M. DEL P. **Metodología de la Investigación**. 6ta. ed. Mexico, DF: McGRAW-HILL / INTERAMERICANA EDITORES, S.A, 2014.

HILL, C. **On Biostatistics and Clinical Trials**. Disponível em: <http://onbiostatistics.blogspot.com.br/2012/05/logx1-data-transformation.html>. Acesso em: 1 jan. 2017.

IHAKA, ROSS ; GENTLEMAN, R. R: a language for data analysis and graphics. **Journal of computational and graphical statistics**, v. 5, n. 3, p. 299--314, 1996.

JAIN, M. et al. Validation of housekeeping genes as internal control for studying gene expression in rice by quantitative real-time PCR. **Biochemical and Biophysical Research Communications**, v. 345, n. 2, p. 646–651, 2006.

KNIGHT, J. et al. Detecting Multivariate Gene Interactions in RNA-Seq Data Using Optimal Bayesian Classification. **IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM**, v. 5963, n. c, p. Epub ahead of print., 2015.

KOGENARU, S. et al. RNA-seq and microarray complement each other in transcriptome profiling. **BMC genomics**, v. 13, n. 1, p. 629, 2012.

KUKURBA, K. R.; MONTGOMERY, S. B. RNA Sequencing and Analysis. **Cold Spring Harbor protocols**, p. pdb.top084970-, 2015.

LERCHER, M. J.; URRUTIA, A. O.; HURST, L. D. Clustering of housekeeping genes provides a unified model of gene order in the human genome. **Nature Genetics**, v. 31, n. 2, p. 180–183, 2002.

LIU, Y. et al. Understanding of intenal clustering validation measures. **IEEE Internatinal Conference on Data mining**, p. 911–916, 2010.

MANGANELLI, R. et al. The Mycobacterium tuberculosis ECF sigma factor sigmaE: role in global gene expression and survival in macrophages. **Molecular microbiology**, v. 41, n. 2, p. 423–37, jul. 2001.

MARTIN, J. A.; WANG, Z. Next-generation transcriptome assembly. **Nature Reviews Genetics**, v. 12, n. 10, p. 671–682, 2011.

MEDINI, D. et al. The microbial pan-genome. **Current Opinion in Genetics and Development**, v. 15, n. 6, p. 589–594, 2005.

MOROZOVA, O.; MARRA, M. A. Genomics Applications of next-generation sequencing technologies in functional genomics. **Genomics**, v. 92, p. 255–264, 2008.

ORACLE. **Oracle® Data Mining Concept**. [s.l.] Oracle, 2008. v. 1

PEPIN, M.; BOISRAME, A.; MARLY, J. Corynebacterium pseudotuberculosis: biochemical properties, production of toxin and virulence of ovine and caprine strains. **Ann Rech Vet**, v. 20, n. 1, p. 111–115, 1989.

PIERCE, B. A. **Genética: Un enfoque conceptual**. [s.l.] Ed. Médica Panamericana, 2009.

PINTO, A. C. et al. Differential transcriptional profile of Corynebacterium pseudotuberculosis in response to abiotic stresses. **BMC genomics**, v. 15, p. 14, 2014.

PINTO, F. et al. Selection of suitable reference genes for RT-qPCR analyses in cyanobacteria. **PLoS ONE**, v. 7, n. 4, p. 1–9, 2012.

POLANSKI, ANDRZEJ ; KIMMEL, M. **Bioinformatics**. 1. ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007.

PUTZER, H.; LAALAMI, S. Regulation of the Expression of Aminoacyl-tRNA Synthetases and Translation Factors. In: **Madame Curie Bioscience Database [Internet]**. Austin (TX): Landes Bioscience, 2000.

QUAIL, M. et al. A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. **BMC Genomics**, v. 13, n. 1, p. 1, 2012.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**VienaR Foundation for Statistical Computing, , 2016. Disponível em: <https://www.r-project.org/>

RAMACHANDRAN, P. et al. Development of a tiered multilocus sequence typing scheme for members of the Lactobacillus acidophilus complex. **Applied and Environmental Microbiology**, v. 79, n. 23, p. 7220–7228, 2013.

RAMAN, S. et al. The Alternative Sigma Factor SigH Regulates Major Components of Oxidative and Heat Stress Responses in Mycobacterium tuberculosis. **Journal of**

Bacteriology, v. 183, n. 20, p. 6119–6125, 15 out. 2001.

RAPAPORT, F. et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. **Genome Biology**, v. 14, n. 9, p. R95, 2013.

RENDÓN, E. et al. Internal versus External cluster validation indexes. **International Journal of Computers and Communications**, v. 5, n. 1, p. 27--34, 2011.

REUE, K.; GLUECK, S. B. Editorial Focus. n. 1, p. 1–2, 2001.

ROCHA, D. J. P.; SANTOS, C. S.; PACHECO, L. G. C. Bacterial reference genes for gene expression studies by RT-qPCR: survey and analysis. **Antonie van Leeuwenhoek**, v. 108, n. 3, p. 685–693, 2015a.

ROCHA, D. J. P.; SANTOS, C. S.; PACHECO, L. G. C. Bacterial reference genes for gene expression studies by RT-qPCR: survey and analysis. **Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology**, v. 108, n. 3, p. 685–693, 2015b.

RODRÍGUEZ, D.; CUADRADO, J.; SICILIA, M. Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software. **del Software**, 2007.

ROKACH, L. A survey of Clustering Algorithms. In: **Data Mining and Knowledge Discovery Handbook**. Boston, MA: Springer US, 2009. p. 269–298.

ROKACH, L. A survey of Clustering Algorithms. In: MAIMON, O.; ROKACH, L. (Eds.). . **Data Mining and Knowledge Discovery Handbook**. Boston, MA: Springer US, 2010. p. 269–298.

ROKACH, L.; MAIMON, O. Chapter 15— Clustering methods. **The Data Mining and Knowledge Discovery Handbook**, p. 32, 2010.

RUIZ, J. C. et al. Evidence for reductive genome evolution and lateral acquisition of virulence functions in two corynebacterium pseudotuberculosis strains. **PLoS ONE**, v. 6, n. 4, 2011.

SELIM, S. A. et al. Immunological characterization of diphtheria toxin recovered from Corynebacterium pseudotuberculosis. **Saudi Journal of Biological Sciences**, v. 23, n. 2, p. 282–287, 2016.

SHENDURE, J.; JI, H. Next-generation DNA sequencing. **Nat Biotechnol**, v. 26, n. 10, p. 1135–1145, 2008.

SI, Y. et al. Model-based clustering for RNA-seq data. **Bioinformatics**, v. 30, n. 2, p. 197–

205, 2014.

SIHTO, H. M. et al. Validation of reference genes for normalization of qPCR mRNA expression levels in Staphylococcus aureus exposed to osmotic and lactic acid stress conditions encountered during food production and preservation. **FEMS Microbiology Letters**, v. 356, n. 1, p. 134–140, 2014.

SNUSTAND, P.; SIMMONS, M. J. **Fundamentos de Genética**. 6ta. ed. Rio de Janeiro: Guanabara Koogan, 2013.

SOARES, S. C. et al. Genome sequence of Corynebacterium pseudotuberculosis biovar equi strain 258 and prediction of antigenic targets to improve biotechnological vaccine production. **Journal of Biotechnology**, v. 167, n. 2, p. 135–141, 2013a.

SOARES, S. C. et al. The Pan-Genome of the Animal Pathogen Corynebacterium pseudotuberculosis Reveals Differences in Genome Plasticity between the Biovar ovis and equi Strains. **PLoS ONE**, v. 8, n. 1, 2013b.

STEPHENS, Z. D. et al. Big data: Astronomical or genomical? **PLoS Biology**, v. 13, n. 7, p. 1–11, 2015.

SZEKERES, E. et al. Selection of proper reference genes for the cyanobacterium Synechococcus PCC 7002 using real-time quantitative PCR. **FEMS Microbiology Letters**, v. 359, n. 1, p. 102–109, 2014.

TETTELIN, H. et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial “pan-genome”. **Proceedings of the National Academy of Sciences of the United States of America**, v. 102, n. 39, p. 13950–5, 2005.

THE GENE ONTOLOGY CONSORTIUM. Expansion of the Gene Ontology knowledgebase and resources. **Nucleic Acids Research**, v. 45, n. November 2016, p. 1–8, 2016.

THELLIN, O. et al. Housekeeping genes as internal standards: use and limits. **Journal of Biotechnology**, v. 75, n. 2–3, p. 291–295, 1999.

THOMAS, F.; BARBEYRON, T.; MICHEL, G. Evaluation of reference genes for real-time quantitative PCR in the marine flavobacterium Zobellia galactanivorans. **Journal of Microbiological Methods**, v. 84, n. 1, p. 61–66, 2011.

WAGNER, G. P.; KIN, K.; LYNCH, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. **Theory in Biosciences**, v. 131, n. 4, p. 281–285, 2012.

WANG, Z.; GERSTEIN, M.; SNYDER, M. RNA-Seq: a revolutionary tool for transcriptomics. **Nature reviews. Genetics**, v. 10, n. 1, p. 57–63, 2009.

WEIGEL, C. **schaechter**. Disponível em: <http://schaechter.asmblog.org/schaechter/2014/06/terms-of-biology-the-pan-genome.html>. Acesso em: 1 jan. 2016.

WITTEN, I. H. et al. **Weka : Practical Machine Learning Tools and Techniques with Java ImplementationsSeminar**, 1999. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.157.9488&rep=rep1&type=pdf>

ZHAO, Y. et al. PGAP: Pan-genomes analysis pipeline. **Bioinformatics**, v. 28, n. 3, p. 416–418, 2012.

ZYPRYCH-WALCZAK, J. et al. The Impact of Normalization Methods on RNA-Seq Data Analysis. **BioMed Research International**, v. 2015, 2015.