Universidade Federal de Santa Catarina
Centro de Ciências Biológicas
Departamento de Microbiologia, Imunologia e Parasitologia
Curso de Licenciatura em Ciências Biológicas

Anelize Baranzeli

Coevolução entre micobacteriófagos e micobactérias

Trabalho de conclusão de curso apresentado ao Curso de Licenciatura em Ciências Biológicas, Centro de Ciências Biológicas da Universidade Federal de Santa Catarina, como requisito parcial para obtenção do título de Licenciado em Ciências Biológicas, sob orientação do Profº. Drº. Daniel Santos Mansur

Florianópolis, 2021

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Baranzeli, Anelize
Coevolução entre micobacteriófagos e micobactérias /
Anelize Baranzeli ; orientador, Daniel Santos Mansur, 2021.
57 p.

Trabalho de Conclusão de Curso (graduação) -
Universidade Federal de Santa Catarina, Centro de Ciências
Biológicas, Graduação em Ciências Biológicas, Florianópolis,
2021.

Inclui referências.

1. Ciências Biológicas. 2. Micobactérias. 3. Fagos. 4.
Coevolução. I. Mansur, Daniel Santos. II. Universidade
Federal de Santa Catarina. Graduação em Ciências Biológicas.
III. Título.

Coevolução entre micobacteriófagos e micobactérias

Este de Trabalho Conclusão de Curso foi julgado adequado para obtenção do Título de Licenciada em Ciências Biológicas, e aprovado em sua forma final pela Banca Examinadora.

Florianópolis, 17 de setembro de 2021.

_____

Profº. Drº. Carlos Roberto Zanetti
Coordenador do Curso

Banca Examinadora:

_____

Profº. Drº. Daniel Santos Mansur
Presidente da banca
Universidade Federal de Santa Catarina

_____

Profº. Drº. André Luiz Barbosa Báfica
Membro Titular
Universidade Federal de Santa Catarina

Profº. Drº. Glauber Wagner
Membro Titular
Universidade Federal de Santa Catarina

Profº. Drº. Guilherme de Toledo e Silva
Membro Suplente
Universidade Federal de Santa Catarina

AGRADECIMENTOS

Gostaria de agradecer primeiramente aos meus pais, Elisa e Fernando, pelo apoio e compreensão.

Ao meu orientador, Profº. Drº. Daniel Santos Mansur, por ter me aceito como estudante de iniciação científica no Lidi, pela ótima orientação, reuniões e discussões de resultados.

Ao Drº. Edgar Kozlova por ter guiado os meus primeiros passos na bioinformática e por sempre estar disponível a ajudar.

A Drª. Gabriela Flavia Rodrigues Luiz por todo auxílio nos experimentos, reuniões de discussão de resultados e por sempre estar disposta a ajudar.

A todos que fazem ou fizeram parte da minha trajetória no Lidi, lugar onde comecei a me envolver com ciência, algo que sempre foi um sonho, um objetivo.

Aos professores Drº. André Báfica, Drº. Glauber Wagner e Drº. Guilherme de Toledo e Silva por terem aceito participar da banca deste trabalho, seja como membro titular ou suplente.

A CAPES, por ter concedido minha bolsa de iniciação científica.

A UFSC, por ter me propiciado uma educação pública, gratuita e de qualidade.

A todos os docentes e demais servidores do CCB por terem feito parte da minha formação, sendo essenciais para que ela acontecesse.

As minhas amigas Ana, Bruna e Vanessa e ao meu amigo William, sem o apoio de vocês e todos os nossos trabalhos em grupo, eu não teria chegado até aqui.

Gostaria de agradecer também aqueles que acompanham minha trajetória desde o ensino médio, Beatriz, Eduardo, Luana, Michelle, Paulo e Tayara.

E a todos que além desses fizeram parte do meu caminho na universidade.

RESUMO

Micobactérias são pertencentes à Ordem Actinomycetales, Subordem Corynebacterineae e Família Mycobacteriaceae, grande parte saprófita de vida livre. No entanto, algumas delas podem ser patogênicas, potencialmente ou raramente patogênicas. Um exemplo importante de micobactéria patogênica é a **Mycobacterium tuberculosis**, causadora da tuberculose. O Brasil se encontra entre os 20 países com mais casos no mundo. Os micobacteriófagos são vírus que infectam as micobactérias, sendo organismos extremamente diversos. Muitos bacteriófagos transmitem genes de virulência para as bactérias hospedeiras, o que caracteriza uma relação de coevolução. O presente trabalho utilizou métodos computacionais para a organização, análise e compreensão dos dados das sequências genômicas para elucidar os possíveis eventos de coevolução entre os organismos. Utilizando 39 sequências de genomas completos de micobactérias e 512 genomas completos de micobacteriofagos, foram identificadas 4.593 regiões de alinhamento, sendo eles encontrados dentro de todas os genomas de micobactérias utilizadas. A análise de 100 pares de base a montante e a jusante das inserções utilizando o CRISPRFinder sugere que essas regiões não são componentes de um sistema CRISPR de micobactérias. Para analisar se os insertos estavam inseridos em ORFs de micobactérias, 64 sequências de proteínas de micobactérias foram alinhadas com os insertos, dessas, 53 obtiveram alinhamento com as inserções, e de um total de 852 alinhamentos, 534 alcançaram 100% de similaridade. Foi possível agrupar as micobactérias em quatro grupos segundo o total de inserções, e em dois grupos segundo a divisão em relação à presença ou ausência de inserto de micobacteriófago específico. Uma análise filogenética baseada na sequência do gene rpoB das micobactérias não foi capaz de estabelecer relação com a divisão das micobactérias. Não foi possível identificar enriquecimento dentro da divisão dos grupos de micobactérias estabelecidos no trabalho. Analisando o gráfico de dispersão entre tamanho do genoma e quantidade de inserções totais foi possível observar uma possível relação entre as duas variáveis. Testes de normalidade e de correlação foram utilizados para testar essa possível relação, sugerindo que há forte correlação entre tamanho do genoma e quantidade de inserções. Esse estudo sugere que nos grupos observados em micobactérias pela quantidade de inserção total não há preferência para inserção nas proteínas de micobactérias e nem

relação com a filogenia, mas sim a quantidade de inserções totais independente de micobacteriófago específico. Novos trabalhos podem utilizar as ferramentas desenvolvidas para estabelecer relações evolutivas dentro de outros conjuntos de sequências genômicas.
Palavras-chave: micobactérias, fagos, coevolução

ABSTRACT

Mycobacteria belong to the Order Actinomycetales, Suborder Corynebacterineae and Family Mycobacteriaceae, a large part of free living saprophytes. However, some of them can be pathogenic, potentially or rarely pathogenic. An important example of a pathogenic mycobacterium is **Mycobacterium tuberculosis**, which causes tuberculosis. Brazil is among the 20 countries with the most cases in the world. Mycobacteriophages are viruses that infect mycobacteria, being extremely diverse organisms. Many bacteriophages transmit virulence genes to host bacteria, which characterizes a coevolutionary relationship. The present work used computational methods for the organization, analysis and understanding of data from genomic sequences to elucidate possible coevolution events between organisms. Using 39 sequences of complete mycobacterial genomes and 512 complete mycobacteriophage genomes, 4,593 alignment regions were identified, and they are found within all mycobacterial genomes used. Analysis of the upstream and downstream 100 base pair inserts using the CRISPRFinder suggests that these regions are not components of a mycobacterial CRISPR system. To analyze whether the inserts were inserted into mycobacterial ORFs, 64 mycobacterial protein sequences were aligned with the inserts, of which 53 achieved alignment with the inserts, and out of a total of 852 alignments, 534 achieved 100% similarity. It was possible to divide the mycobacteria into four groups according to the total number of insertions, and into two groups according to the division in relation to the presence or absence of a specific mycobacteriophage insert. A phylogenetic analysis based on the sequence of the mycobacteria rpoB gene was not able to establish a relationship with the division of mycobacteria. It was not possible to identify enrichment within the division of the mycobacteria groups established in the work. Analyzing the scatter plot between genome size and number of total insertions it was possible to observe a possible relationship between the two variables and then normality and correlation tests were used to test this possible relationship, the result was that there is a strong correlation between size of the genome and number of insertions. This study suggests that in the groups observed in mycobacteria due to the amount of total insertion, there is no preference for insertion in the mycobacterial proteins and neither is it related to

phylogeny, but the number of total insertions is independent of specific mycobacteriophage. New works can use the tools developed to establish evolutionary relationships within other sets of genomic sequences.
Keywords: mycobacteria, phages, coevolution

LISTA DE FIGURAS

LISTA DE TABELAS

LISTAS DE ABREVIATURAS E SIGLAS

CMTB - Complexo **Mycobacterium tuberculosis**

CRISPR - Clustered Regularly Interspaced Short Palindromic Repeats

Fagos - Bacteriófagos

Gb – Gigabyte

GO - Gene Ontology

HD – Disco rígido

Kb – Mil pares de base

Mb - Milhões de pares de bases

MNT - Micobactérias não tuberculosas

NCBI - National Center for Biotechnology Information

PB - Pares de base

PAM - Partitioning Around Medoids

Pamk - Partitioning Around Medoids With Estimation of Number of Clusters

PCA - Principal Components Analysis

rpoB - subunidade β de RNA polimerase

Tb – Terabyte

SUMÁRIO

1. INTRODUÇÃO

1.1 Micobactérias

Micobactérias são bactérias de formato bacilar pertencentes à Ordem Actinomycetales, Subordem Corynebacterineae e Família Mycobacteriaceae (BRASIL, 2005), a largura varia entre 0.2μm e 0.6μm e o comprimento entre de 1μm e até 10μm. Possuem respiração aeróbia ou microaerófila, tempo de crescimento variado, podendo ser divididas entre micobactérias de crescimento rápido e crescimento lento. Grande parte saprófita, completando seu ciclo no ambiente, mas algumas são intracelulares (Barrera, 2007). Além disso, são consideradas gram-positivas, porém, como possuem uma alta quantidade de lipídeo integrando a parede celular, não são tipicamente coradas (BRASIL, 2009).

As micobactérias são classificadas em dois grandes grupos, o complexo **Mycobacterium tuberculosis** (CMTB), formados pelas espécies **M. tuberculosis**, **M. africanum**, **M. canettii**, **M. microti** e **M. bovis**, que causam a tuberculose e possuem 99.9% de identidade genética entre si (BROSCH et al., 2002) com associações geográficas para cada uma delas (COMAS et al., 2010). A sua origem monofilética e sua expansão em diferentes partes do mundo sugere que poderiam ter acompanhado as primeiras migrações fora da África e diversificado em conjunto com diferentes populações humanas (HERSHBERG et al., 2008, WIRTH et al., 2008). O segundo grupo é composto por micobactérias não tuberculosas (MNT), que são geralmente oportunistas (UEKI, 2005), ou seja, afetam indivíduos que possuem seu sistema imunológico prejudicado (SANTANA; SILVA; PEREIRA, 2019), são saprófitas de vida livre (SANTOS, 2015), podendo ocasionar doenças pulmonares, nas glândulas linfáticas, em feridas, ossos (KATOCH, 2004), pele e tecidos após procedimentos cirúrgicos (PÔSSA, 2011), sendo encontradas na água (COLLINS; GRANGE; YATES, 1984) e solo (FREY, 1930).

Essas bactérias também podem ser classificadas quanto a sua aptidão de causar doenças no homem, sendo patogênicas quando causam necessariamente, potencialmente

patogênicas quando podem causar e raramente patogênicas quando nunca ou raramente causam doenças (BRASIL, 2008).

Uma importante bactéria dentro do grupo é a **Mycobacterium tuberculosis**, uma das causadoras da tuberculose, doença infecciosa que prejudica principalmente os pulmões (PANDOLFI et al., 2007). A estimativa é que um terço da população tenha sido infectada pela doença, e o Brasil está entre os 20 países com mais casos no mundo (BRASIL, 2017).

1.2 Bactériófagos

Especula-se que os bacteriófagos (fagos) sejam os mais diversos organismos do ecossistema, podendo ser considerados a matéria escura da biologia (PEDULLA et al., 2003). No mundo, estima-se que aproximadamente 10^{25} infecções por fago se iniciem por segundo, onde em cada infecção o DNA pode recombinar-se gerando novos arranjos genômicos (HENDRIX, 1978), sendo que muitos transmitem genes de virulência que codificam proteínas com função fundamental na patogênese bacteriana (BOYD; DAVIS; HOCHHUT, 2001). Existem evidências de profagos no genoma de micobactérias, como é o caso da **Mycobacterium abscessus**, onde representam entre 6.7 e 9.6% do seu conteúdo genômico (SASSI et al., 2014).

Os micobacteriófagos fazem parte da ordem Caudovirales e em relação a família, os que possuem caudas não contráteis e moderadamente longas pertencem a família Siphoviridae, já aqueles que têm caudas contráteis pertencem a família Myoviridae (HATFULL, 2018). A maioria dos micobacteriófagos já caracterizados apresentam dupla fita de DNA, possuindo um capsídeo, com tamanho relacionado com a extensão do genoma (HATFULL et al., 2010), sendo considerado um mosaico e podendo ser dividido em ‘clusters’ e ‘subclusters’ com base em semelhanças nos seus genes e onde estes estão codificados no DNA (POPE et al., 2015).

1.3 Genômica

Os genomas do gênero Mycobacterium têm o tamanho variando entre 3.20 e 8 Mb, tendo a média entre 6.6 e 6.7 Mb, contendo entre 57% e 69% de GC (MORGADO, 2017). Já o tamanho do genoma dos micobacteriófagos é entre 41.4 e 164.6 Kb com média de 73.6 Kb (HATFUL, 2010), apresentando entre 57.3% e 69% de GC no seu genoma (PEDULLA et al., 2003).

Alguns fatores corroboram para uma relação de coevolução, situação que ocorre quando as alterações evolutivas no patógeno aumentam a sua capacidade de infecção e são compensadas por alterações evolutivas no hospedeiro que aumentam a resistência à infecção (THOMPSON, 1994), entre todo o gênero Mycobacterium e os micobacteriófagos.

Um primeiro fator é a hipótese da Rainha Vermelha, que diz que o ecossistema dos seres vivos está continuamente se deteriorando, causando nas espécies um ininterrupto empenho para permanecerem adaptadas (VAN VALEN, 1973). Segundo a realização de conversão lisogênica feita pelos bacteriófagos, ou seja, podem transformar uma cepa não patogênica em patogênica ou com virulência aumentada (PETERS et al., 2019, WALDOR; MEKALANOS, 1996). Por último, o fato de que o sucesso ecológico de um ciclo lisogênico contribui para a disseminação de genes de bacteriófagos (DESIERE et al., 2001).

Dada a importância da tuberculose, os estudos com os micobacteriófagos, podem demonstrar utilidade contra infecções bacterianas através da fagoterapia, que é o uso de um bacteriófago para tratar uma infecção bacteriana. O método é considerado desde os anos 80 como uma solução para a resistência das bactérias aos antibióticos (DE FREITAS ALMEIDA; SUNDBERG, 2020).

Para compreender melhor a interação de micobactérias e micobacteriófagos, a análise genômica dos dois organismos por métodos utilizando bioinformática se faz necessária, visto que milhares de pares de bases de sequências de nucleotídeos são inseridas em bancos de dados todos os dias e toda essa quantidade de informação só é

viável de ser ordenada, examinada e compreendida com o suporte da informática (SANTOS; ORTEGA, 2003).

Outro exemplo onde a bioinformática se faz necessária é com as regiões de sequências de nucleotídeos repetitivas em bactérias e arqueas, em inglês Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR), um sistema de imunidade adaptativa das arquéias e bactérias (JANSEN et al., 2002), que embora tenha sua primeira descrição feita em 1987 (ISHINO et al., 1987), só foi nomeada quando o estudo *in silico* se fez presente (JANSEN et al., 2002) e a partir desse momento, amplamente estudada, além de existirem ferramentas de análise de CRISPR totalmente *in silico* , como o CRISPRFinder, que é uma ferramenta online para identificação dessas repetições palindrômicas curtas agrupadas e regularmente interespaçadas em sequências nucleotídicas (GRISSA et al., 2007).

O campo da bioinformática relaciona áreas diferentes para compreender, por exemplo, o genoma através de modelos matemáticos e estatísticos, mas também cria programas computacionais para processamento de dados biológicos, gera prognóstico do aspecto tridimensional das proteínas, determina árvores filogenéticas, entre diversas outras atribuições. (DIAS DE ARAÚJO et al., 2008).

Embora a relação entre micobacteriófagos e micobactérias seja estudada (SASSI et al., 2014; GENTILE et al., 2019), faltam trabalhos que utilizam uma ampla variedade tanto de micobacteriófagos como de micobactérias, incluindo as não patogênicas.

O presente trabalho visa o estudo da coevolução entre micobacteriófagos e micobactérias, a fim de elucidar os possíveis dispositivos de coevolução entre os organismos.

2. OBJETIVOS

2.1 Objetivo Geral

Estudar a relação entre micobacteriófagos e micobactérias, a fim de elucidar os possíveis dispositivos de coevolução entre os organismos.

2.2 Objetivos Específicos

a) Levantar os genomas disponíveis de várias espécies de micobactérias nos bancos de dados públicos.

b) Criar um programa para localizar inserções de genoma de micobacteriófagos no genoma de micobactérias.

c) Analisar, *in silico*, as inserções identificadas dentro do sistema CRISPR de micobactérias e também a sua função biológica entre as proteínas de micobactérias.

d) Classificar as micobactérias em grupos segundo a quantidade de inserções apresentadas.

e) Investigar o motivo pelo qual as micobactérias têm quantidade de inserções diferentes.

3. MATERIAL E MÉTODOS

3.1 Descrição da máquina utilizada

A máquina utilizada no presente trabalho possui processador Intel ® Xeon ® com 48 cores, memória RAM de 128Gb, 8 Tb de HD e sistema operacional Linux.

3.2 Sequências genômicas e proteômicas

A seleção de proteomas e genomas foi feita a partir da busca de genomas e proteomas completos de micobactérias e micobacteriófagos disponíveis no National Center for Biotechnology Information (NCBI) (COORDINATORS, 2016), disponíveis no ano de 2018. A lista com os números de acesso dos genomas que foram utilizados e analisados está na tabela: número de acesso NCBI, o seu download pode ser realizado através: https://github.com/anelizebaranzeli1997/NumeroDeAcessoNCBI.

Para o estudo das posições dos insertos em proteínas de micobactérias foi feito o download da sequência de proteínas de micobactérias no banco de dados Uniprot (BATEMAN, 2019), excluindo as não caracterizadas. A lista com os números de acesso das sequências de proteínas está na tabela intitulada: número de acesso Uniprot, disponível em: https://github.com/anelizebaranzeli1997/NumeroDeAcessoUniprot.

3.3 Alinhamentos

O alinhamento entre as sequências de micobactérias e micobacteriófagos foi feito através do Blastn (ALTSCHUP et al., 1990), utilizando os parâmetros padrão, que compara os nucleotídeos entre as sequências sendo um método de alinhamento local, onde é efetuada uma busca por regiões com semelhança local e não é considerado todo o comprimento da sequência.

O alinhamento entre as inserções e as sequências de proteínas foi realizado através do Blastx, semelhante ao Blastn, porém adequado para o alinhamento entre nucleotídeos e proteínas, utilizando os parâmetros padrão, limitando o e-value em 1e-5 e apenas um resultado por sequência.

Os comandos utilizados podem ser verificados em https://github.com/anelizebaranzeli1997/Comandos-Blast.

3.4 Implementação da ferramenta

Foi escrito um programa na linguagem Python (VAN ROSSUM G., 1995) para processar os dados do alinhamento, selecionando apenas as partes que alinharam, onde após a organização dos dados foi investigado a sua origem no genoma das micobactérias.

Outro programa foi desenvolvido na linguagem Python (VAN ROSSUM G., 1995), o qual estende a sequência analisada de nucleotídeos além da zona de alinhamento entre as micobactérias e os micobacteriófagos.

3.5 CRISPR

A possibilidade desses insertos serem derivados do sistema CRISPR foi averiguada com a utilização do programa CRISPRFinder (GRISSA et al., 2007).

3.6 Análise de função biológica

O resultado do alinhamento entre proteínas e inserções foi processado e utilizado para investigar a função das proteínas as quais estão inseridos os insertos de micobacteriófagos utilizando a plataforma Blast2go (CONESA et al., 2005) que é ideal para anotação funcional e análise de conjuntos de dados genômicos.

3.7 Análises dos agrupamentos

Foi avaliado de que maneira estão agrupadas as micobactérias em relação a quantidade de inserção e fago específico inserido no seu genoma. Para isso, as análises foram feitas utilizando as linguagens bash, R (R CORE TEAM, 2014) e as funções pamk (HENNING, 2018) e prcomp (DUNTEMAN, 1989).

3.8 Análise de enriquecimento

Para as análises de enriquecimento dos insertos posicionados em proteínas de micobactérias por grupo foi utilizada a plataforma ShinyGO v0.66 disponível em http://bioinformatics.sdstate.edu/go/ (GE et al., 2020).

3.9 Cladograma

As sequências do gene subunidade β de RNA polimerase (rpoB) foram obtidas através da busca avançada do NCBI (COORDINATORS, 2016), os números de acesso estão disponíveis em https://github.com/anelizebaranzeli1997/Numero-de-Acesso-rpoB.

Para o alinhamento entre as sequências do gene rpoB o método de alinhamento global Clustal Omega (SIEVERS; HIGGINS, 2014) foi utilizado.

A análise filogenética foi realizada pelo método de probabilidade de clusterização, utilizando o pacote pvclust (SUZUKI; SHIMODAIRA, 2006). A visualização do cladograma se deu através do serviço de web ETE Toolkit (HUERTA-CEPAS; SERRA; BORK, 2016) disponível em http://etetoolkit.org/treeview/, com posterior edição manual.

3.10 Análise da quantidade de inserções por tamanho do genoma

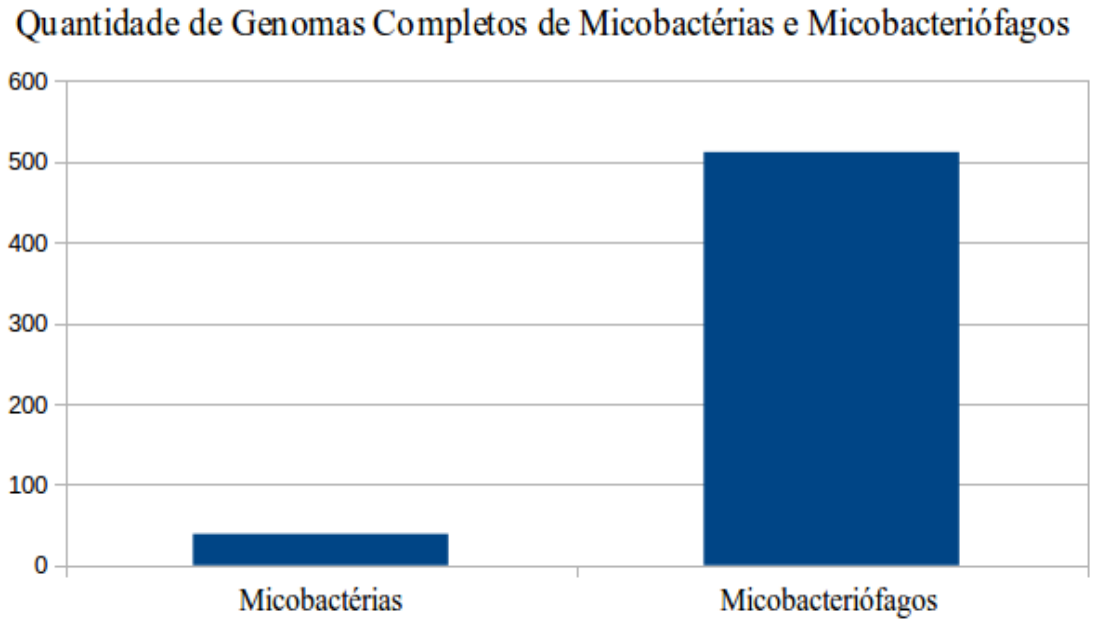
Para o estudo da relação entre o tamanho do genoma das micobactérias e a quantidade de inserções por grupo de bactérias agrupadas pela quantidade de inserção de fagos específicos foi utilizado o pacote ggplot2 (VALERO-MORA, 2010) do R (R CORE TEAM, 2014), com posterior edição manual no gráfico de dispersão. Para as demais análises utilizou-se o teste de Shapiro-Wilk (SHAPIRO; WILK, 1965) e coeficiente de correlação de Spearman (ZAR, 2005).

4. RESULTADOS

4.1 Inserções de genoma de micobacteriófagos no genoma de micobactérias

Para verificar as inserções de micobacteriófagos no genoma das micobactérias foi realizado o alinhamento das sequências de micobactérias e micobacteriófagos através do Blastn, utilizando a quantidade de genomas completos que pode ser visto na figura a seguir.

FIGURA 1 - Comparação entre a quantidade de genomas completos de micobactérias e micobacteriófagos utilizados neste trabalho.



Fonte: Elaborada pela autora, 2021.

Obtevê-se um grande volume de dados que podem ser conferidos através do link: https://github.com/anelizebaranzeli1997/Tabela-de-resultado-de-alinhamentos-entre-micobacterias-e-micobacteriofagos.

Abaixo pode-se observar uma amostra dos resultados obtidos no alinhamento.

TABELA 1 - Amostra do resultado de alinhamento blastn entre micobactérias e micobacteriófagos.

Identificador da sequência de interesse	Identificador da sequência do banco de dados	% de identidade	Tamanho do alinhamento	PB diferentes	Aberturas	Posição inicial na sequência de interesse	Posição final na sequência de interesse	Posição inicial na sequência do banco de dados	Posição final na sequência do banco de dados	E value	Bit score
CH482373.1	NC_023580.1	100.00	28	0	0	775012	775039	22730	22703	0,00002	52.8
CP001664.1	NC_021334.1	100.00	28	0	0	775012	775039	31867	31894	0,00002	52.8
CP001664.1	NC_023580.1	100.00	28	0	0	1118521	1118548	22730	22703	0,00003	52.8
CP010333.1	NC_021334.1	100.00	28	0	0	1118521	1118548	31867	31894	0,00003	52.8
CP010333.1	NC_023580.1	100.00	28	0	0	1118114	1118141	22730	22703	0,00003	52.8
CP015773.2	NC_021334.1	100.00	28	0	0	1118114	1118141	31867	31894	0,00003	52.8
CP015773.2	NC_023580.1	100.00	28	0	0	1114491	1114518	22730	22703	0,00003	52.8
CP015773.2	NC_023580.1	100.00	28	0	0	1114491	1114518	31867	31894	0,00003	52.8
CP021238.1	NC_026588.1	100.00	28	0	0	3233372	3233399	28858	28831	0,00003	52.8

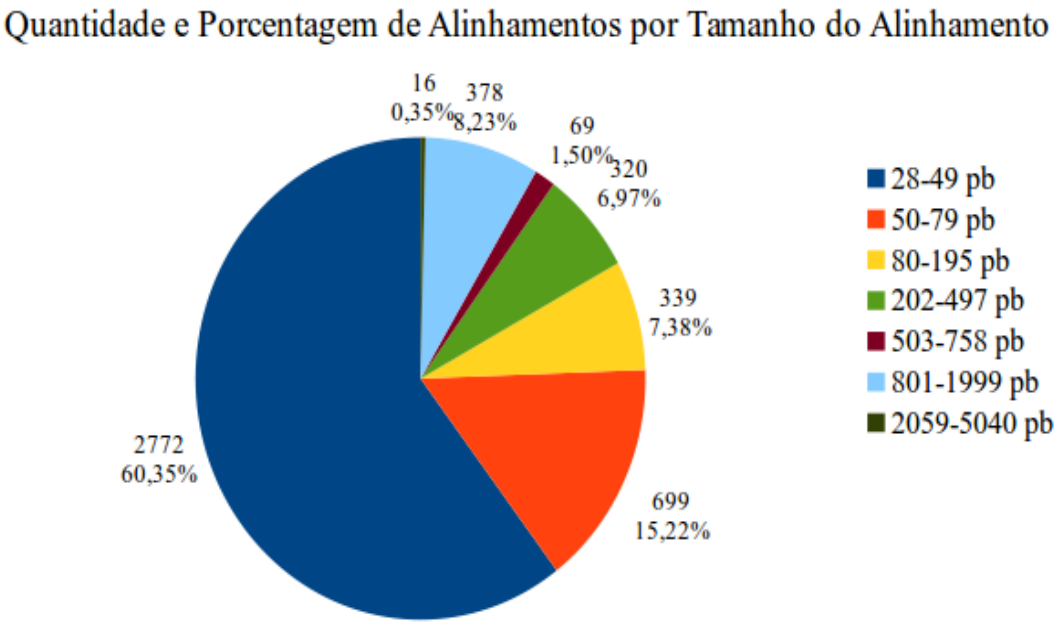
Fonte: Elaborada pela autora, 2021

Um baixo valor de e value significa uma melhor qualidade do alinhamento, pois ele descreve o número de ocorrências que se espera ver por acaso ao pesquisar em um banco de dados de um tamanho específico, por isso quanto menor o seu valor, ou mais próximo de zero, mais significativo é o alinhamento (NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION (US); CAMACHO, 2008).

Já o bit score, é a versão em escala de log do score, que é um número usado para avaliar a relevância biológica de um achado, e dentro do contexto de alinhamentos, números mais altos correspondem a semelhanças mais altas (NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION (US); CAMACHO, 2008).

Houve alinhamento em todos os genomas de micobactérias, com o tamanho do alinhamento variando entre 28 a 5040 pares de base, a distribuição em quantidade e porcentagem dos tamanhos do alinhamento encontrados pode ser observada no gráfico abaixo.

FIGURA 2 - Gráfico demonstrando a variação em quantidade e porcentagem de alinhamentos por tamanho do alinhamento.



Fonte: Elaborada pela autora, 2021.

4.2 Desenvolvimento das ferramentas

Para analisar separadamente as inserções foi criado o primeiro programa, o getPosition.py, utilizado para obter as partes específicas alinhadas entre micobactérias e micobacteriófagos está disponível para download e uso através do link: https://github.com/anelizebaranzeli1997/getPositionScript.

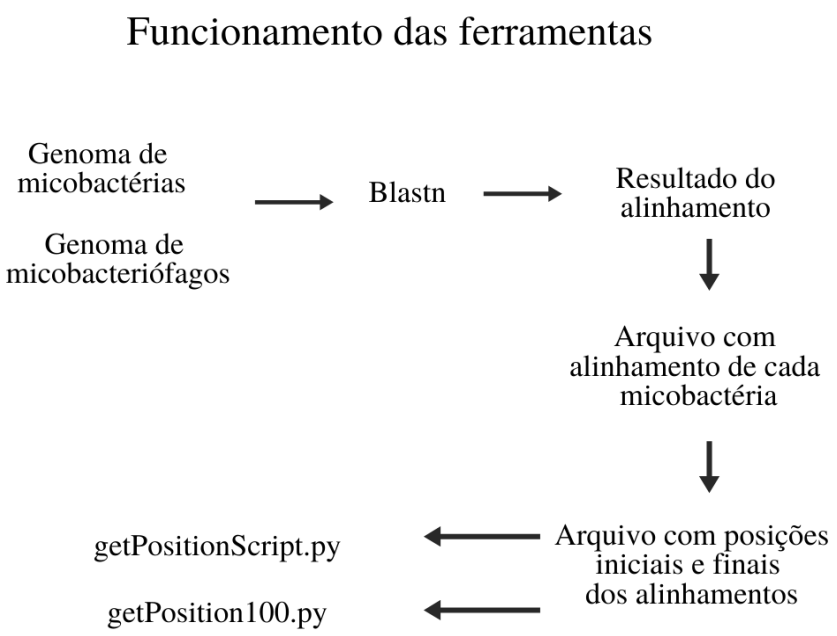
Para utilizá-lo, primeiro foram retirados os alinhamentos de cada micobactéria do alinhamento geral entre micobacteriófagos e micobactérias utilizando comandos do bash, depois separa-se as posições iniciais e finais de cada alinhamento por cada micobactéria do alinhamento geral entre micobacteriófagos e micobactérias. Estando no diretório correto onde os arquivos necessários estão armazenados executa se o comando: python getPositionScript.py ArquivoComPosições SequênciaDaMicobactériaDeInteresse.fasta ArquivoDeSaída.

Para o estudo de CRISPR, um segundo programa foi criado, o getPosition100.py, para obter as partes específicas alinhadas entre micobactérias e micobacteriófagos, com uma extensão de 100 nucleotídeos a direita e à esquerda da parte específica alinhada, que está disponível para download e uso através do link: https://github.com/anelizebaranzeli1997/getPosition100.

Para utilizá-lo, é necessário ter um arquivo com a posição inicial e posição final de cada alinhamento entre micobactérias e micobacteriófagos, seu uso se dá por linha de comando. Estando no diretório correto, onde os arquivos necessários estão armazenados através do comando: python getPosition100.py ArquivoComPosições SequênciaDaMicobactériaDeInteresse.fasta ArquivoDeSaída.

Na imagem abaixo é possível observar um fluxograma de funcionamento das duas ferramentas.

FIGURA 3 - Fluxograma do funcionamento das ferramentas criadas e utilizadas no trabalho.



Fonte: Elaborada pela autora, 2021.

4.3 Análise CRISPR

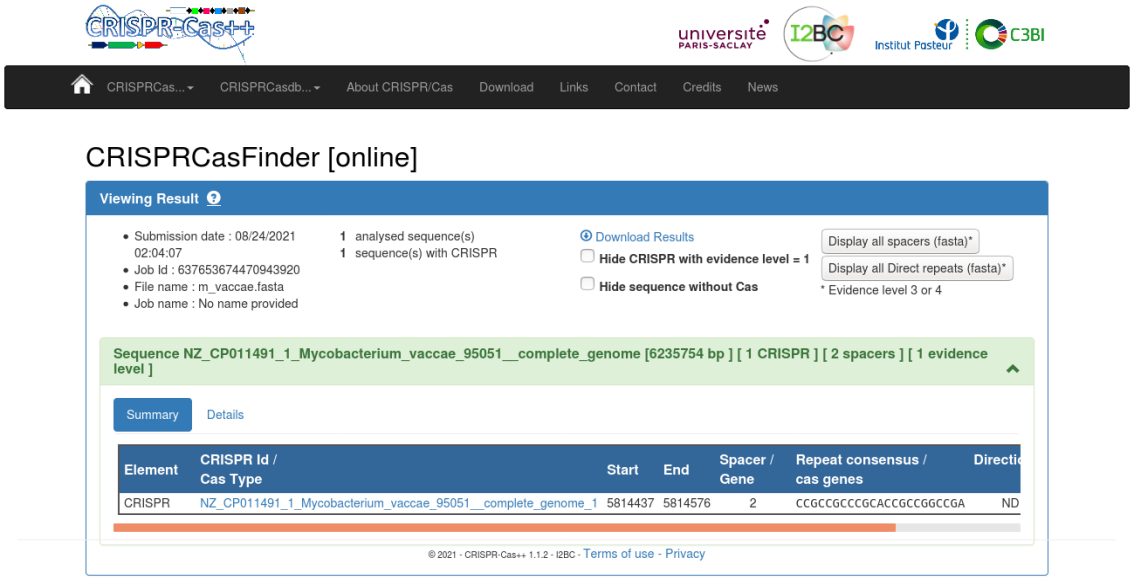
Sendo parte dos insertos compatíveis com o tamanho de espaçadores considerados pelo CRISPRFinder, entre 25 e 60 pares de base (GRISSA et al., 2007), a análise para verificar se estariam localizados em regiões de CRISPR de micobactérias foi realizada. Os insertos foram adicionados de 100 pares de base através do programa getPosition100.py. As sequências podem ser visualizadas através do link https://github.com/anelizebaranzeli1997/-anelizebaranzeli1997-Insertos-adicionados-100-pares-de-base-, elas foram utilizadas para realizar uma análise utilizando o CRISPRFinder (GRISSA et al., 2007), um programa que identifica possíveis áreas de CRISPR em sequências nucleotídicas.

O programa busca o tamanho máximo de repetições, entre 23 a 55 pares de base, espaçados por sequências entre 25 a 60 pares de base. As regiões conservadas no genoma são selecionadas com a repetição máxima que ocorre mais vezes em todo o genoma. Tendo as regiões conservadas no genoma determinadas, os espaçadores são

extraídos, descartando-se as repetições em tandem comparando espaçador com espaçador ou região conservada no genoma com espaçador. Os candidatos com mais de três espaçadores são considerados CRISPRs confirmados, com menos espaçadores, são considerados possíveis CRISPRs (GRISSA et al., 2007).

Na figura abaixo temos um exemplo de como os resultados são mostrados quando se utiliza o site.

FIGURA 4 - *Print screen* do exemplo de resultado obtido.



Fonte: *Print screen* de um resultado elaborado pela autora utilizando o CrisprFinder (GRISSA et al., 2007), 2021.

Os resultados que retornaram possíveis áreas de CRISPR nas micobactérias analisadas se encontram em https://github.com/anelizebaranzeli1997/Possiveis-CRISPRs .

As sequências adicionadas de 100 pares de base e os possíveis CRISPR foram alinhados através do Blastn (ALTSCHUP et al., 1990) e o resultado pode ser visualizado com o link a seguir: https://github.com/anelizebaranzeli1997/-Resultado-alinhamento-entre-insertos-adicionados-100-pares-de-base-e-possiveis-CRISPRs . Nenhuma inserção adicionada de 100

pares de base foi detectada como fazendo parte de um possível CRISPR de micobactéria.

4.4 Insertos posicionados em proteínas de micobactérias

Para estudar a possível relação dos insertos nas proteínas de micobactérias, as sequências dos insertos obtidas com o programa getPosition.py, disponíveis em https://github.com/anelizebaranzeli1997/Insertos, foram alinhados com proteínas de micobacteriófagos. O alinhamento gerou uma grande quantidade de dados e por isso está disponível através do link a seguir: https://github.com/anelizebaranzeli1997/Insertos-posicionados-em-proteinas-de-micobacterias.

Os resultados com porcentagem de alinhamento menor do que 100% foram retirados, e então a tabela disponível em https://github.com/anelizebaranzeli1997/Funcao-das-proteinas-dos-insertos-posicionados-em-proteinas-de-micobacterias foi construída, onde pode-se observar qual inserção alinhou com qual proteína, além do tamanho do alinhamento, tamanho e nome da proteína e o link Uniprot (BATEMAN, 2019). Uma amostra dela pode ser observada abaixo.

TABELA 2 – Amostra da tabela de insertos posicionados em proteínas de micobactérias.

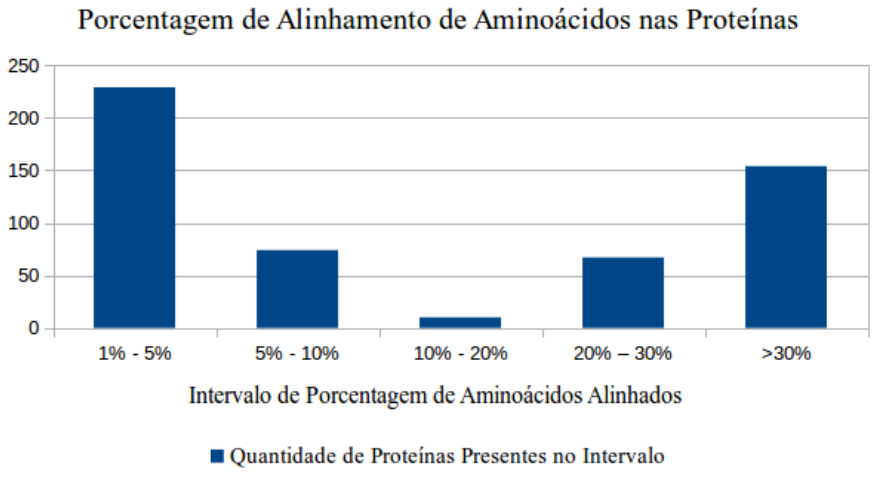
Inserção	Código Uniprot	% de identidade	Tamanho do alinhamento	Tamanho da proteína	Nome da proteína	Link Uniprot
m_sp._JLS_31	A0A1X0GLF5	100.00	16	496	Terminase_6 domain-containing protein	https://www.uniprot.org/uniprot/A0A1X0GLF5
m_sp._JLS_32	A0A1X0GLF5	100.00	16	496	Terminase_6 domain-containing protein	https://www.uniprot.org/uniprot/A0A1X0GLF5
m_sp._JLS_10	A0A1X0GLF5	100.00	16	496	Terminase_6 domain-containing protein	https://www.uniprot.org/uniprot/A0A1X0GLF5
m_sp._JLS_11	A0A1X0GLF5	100.00	16	496	Terminase_6 domain-containing	https://www.uniprot.org/uniprot/A0A1

					protein	X0GLF5
m_sp._JLS_129	A0A1X0GLF5	100.00	17	496	Terminase_6 domain-containing protein	https://www.uniprot.org/uniprot/A0A1X0GLF5
m_sp._JLS_12	A0A1X0GLF5	100.00	16	496	Terminase_6 domain-containing protein	https://www.uniprot.org/uniprot/A0A1X0GLF5
m_sp._JLS_130	A0A1X0GLF5	100.00	17	496	Terminase_6 domain-containing protein	https://www.uniprot.org/uniprot/A0A1X0GLF5
m_sp._JLS_131	A0A1X0GLF5	100.00	17	496	Terminase_6 domain-containing protein	https://www.uniprot.org/uniprot/A0A1X0GLF5
m_sp._JLS_132	A0A1X0GLF5	100	17	496	Terminase_6 domain-containing protein	https://www.uniprot.org/uniprot/A0A1X0GLF5

Fonte: Elaborada pela autora, 2021

Há insertos posicionados em proteínas de micobactérias, cobrindo diferentes porcentagens dos seus proteomas, e para uma melhor visualização desses dados, foram contabilizados a quantidade de proteínas presentes em intervalos específicos de porcentagem de aminoácidos alinhados, o que está disponível em https://github.com/anelizebaranzeli1997/Intervalo-X-Quantidade e pode ser visualizado no gráfico a seguir.

FIGURA 5 - Gráfico que representa a porcentagem de alinhamento de aminoácidos nas proteínas com a quantidade de proteínas presentes em cada um dos intervalos.



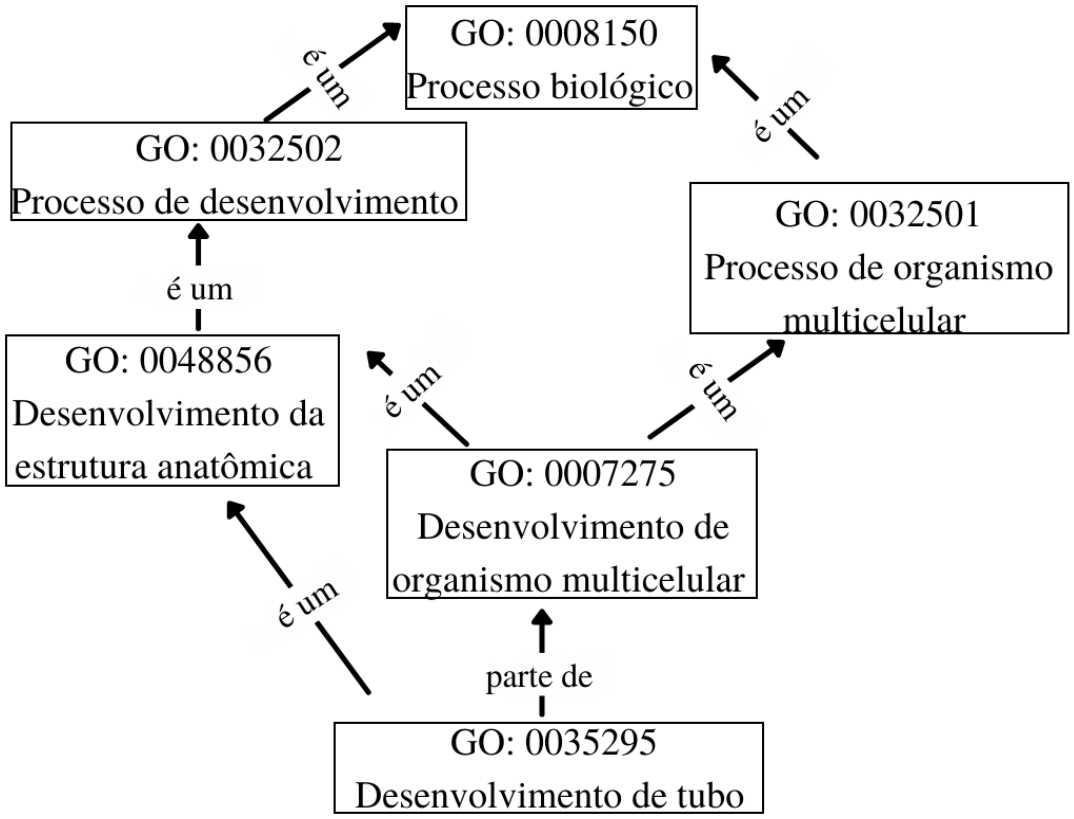
Fonte: Elaborada pela autora, 2021

4.5 Análise da função biológica das proteínas de micobactérias com inserto

Selecionando-se todas as proteínas de micobactérias que alinharam com insertos de micobacteriófagos em micobactérias foi realizada uma análise de enriquecimento para termos do Gene Ontology (GO) utilizando o Blast2go, que é uma ferramenta que permite a análise da associação das sequências proteicas com sua respectiva atribuição biológica utilizando o banco de dados de termos do GO (CONESA et al., 2005).

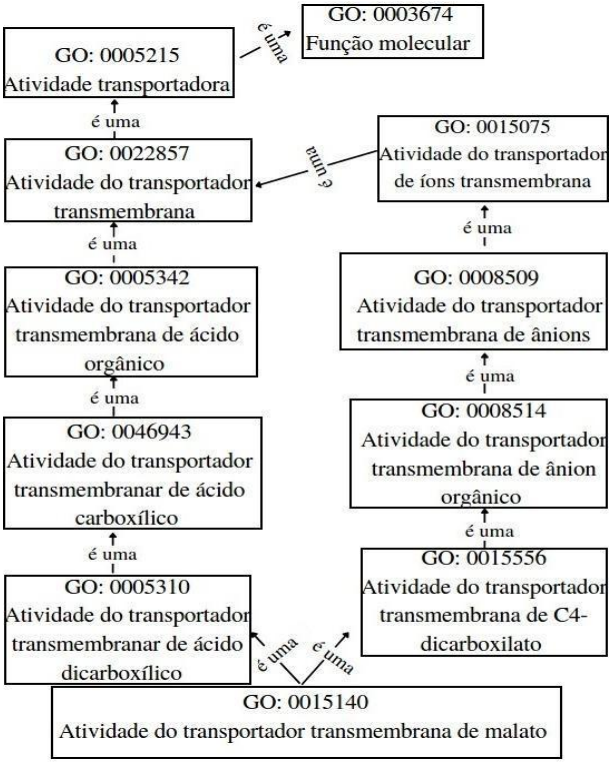
Nos dois grafos a seguir é possível observar quais foram os resultados obtidos com essa análise. As imagens foram modificadas e os grafos originais encontram-se em https://github.com/anelizebaranzeli1997/Grafos-Blast2go.

FIGURA 6 - Grafo modificado com o resultado de processo biológico obtido utilizando o Blast2go (CONESA et al., 2005).



Fonte: Elaborada pela autora baseado no grafo obtido no Blast2go (CONESA et al., 2005), 2021.

FIGURA 7 - Grafo modificado com o resultado de função molecular obtido utilizando o Blast2go (CONESA et al., 2005).



Fonte: Elaborada pela autora baseado no grafo obtido no Blast2go (CONESA et al., 2005), 2021.

Como não houve enriquecimento para termos do GO (Blast2go), não foi possível estabelecer uma associação entre as proteínas com inserto e alguma via ou termo específico.

4.6 Análise de presença ou ausência de micobacteriófago específico e quantidade de inserção de micobacteriófago nas micobactérias.

Para verificar se seria possível agrupar as micobactérias segundo o número de inserções totais de micobacteriófagos foi criado um programa em bash disponível em https://github.com/anelizebaranzeli1997/contagem.sh.

O programa processa os arquivos de entrada sendo um referente aos nomes das linhas (https://github.com/anelizebaranzeli1997/Linhas), outro às colunas (https://github.com/anelizebaranzeli1997/Colunas) e o último, a informação dos

alinhamentos(https://github.com/anelizebaranzeli1997/Info-alinhamento-entre-bacterias-e-fagos-sem-duplicatas-).

Micobacteriófagos que possuem números de referência diferentes, ou seja, mais de uma sequência para a mesma espécie, foram escolhidos para ficarem com apenas um número de referência, pois de outra forma cada inserção seria contada mais de uma vez na micobactéria pelo mesmo micobacteriófago, visto que os resultados de alinhamentos de micobacteriófagos iguais com número de referência diferentes eram os mesmos. Tais arquivos foram gerados a partir do resultado de alinhamento entre sequências de micobactérias e micobacteriófagos utilizando comandos em bash.

Esse programa gerou um arquivo de saída que pode ser verificado em https://github.com/anelizebaranzeli1997/Arquivo-saida-programa-de-contagem, e para a construção da tabela de contagens de inserção de micobacteriófago por bactéria, usa-se as linhas de comando disponíveis em https://github.com/anelizebaranzeli1997/Linhas-de-comando-em-R-para-gerar-tabela-de-contagem. O resultado pode ser visualizado em https://github.com/anelizebaranzeli1997/Tabela-contagem-de-insercao-de-fago-por-bacteria.

E para testar se existe algum micobacteriófago específico que tenha mais inserções em algum dos grupos foi criado uma tabela de presença ou ausência de inserção de micobacteriófago em micobactéria, a tabela de contagens de inserções foi utilizada como base, onde a ausência é representada como 0 e a presença como 1. O resultado está disponibilizado pelo link: https://github.com/anelizebaranzeli1997/Tabela-presenca-ou-ausencia-de-insercao-de-fago-por-bacteria .

A construção de ambas as tabelas foi possível, e elas foram desenvolvidas para serem utilizadas nas análises a seguir.

4.7 Análise dos agrupamentos baseados na presença ou ausência de micobacteriófago específico e quantidade de inserção de micobacteriófago nas bactérias.

Para analisar a partir das tabelas de presença ou ausência de micobacteriófago específico e quantidade de inserção de micobacteriófago nas micobactérias a distribuição em grupo devido aos dois parâmetros, e se a divisão pela quantidade se dava por presença ou ausência de micobacteriófago específico, utilizou-se o método Partitioning Around Medoids With Estimation of Number of Clusters (pamk) (HENNIG, 2015).

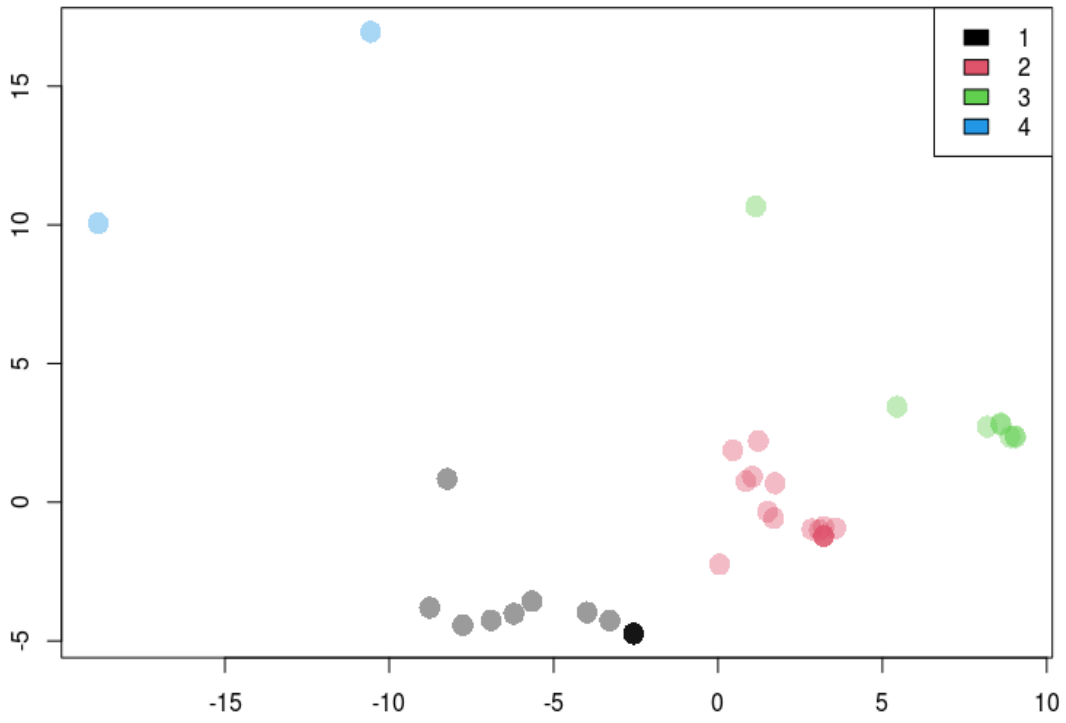
O pamk utiliza o método Partitioning Around Medoids (PAM) (KAUFMAN; ROUSSEEUW, 1990) para definir os grupos, fazendo uma estimação do número desses, sendo PAM um algoritmo que aleatoriamente escolhe os *k* objetos dentro dos dados disponibilizados, sendo esses os primeiros centros, chamados de centróide dos *k* grupos, e então cada objeto é designado ao grupo com centróide mais similar. Esse processo é repetido até que não haja mais mudanças nos centros dos grupos (FABIENE, 2009).

E para uma melhor visualização dos resultados obtidos, foi utilizado o método Principal Components Analysis (PCA), algoritmo matemático usado para reduzir a dimensionalidade dos dados e facilitar a visualização, o que é feito construindo-se combinações lineares com as variáveis originais, onde tais combinações lineares são os componentes principais, podendo dessa maneira identificar padrões ocultos no conjunto de dados. (PEARSON, 1901), que podem ser conferidos em https://github.com/anelizebaranzeli1997/Comandos-PCA.

No gráfico, cada cor corresponde a um grupo observado no resultado obtido através do pamk, sendo possível observar essa relação na legenda das figuras 8 e 9. Quanto mais pontos sobrepostos, menor é a transparência do ponto no local.

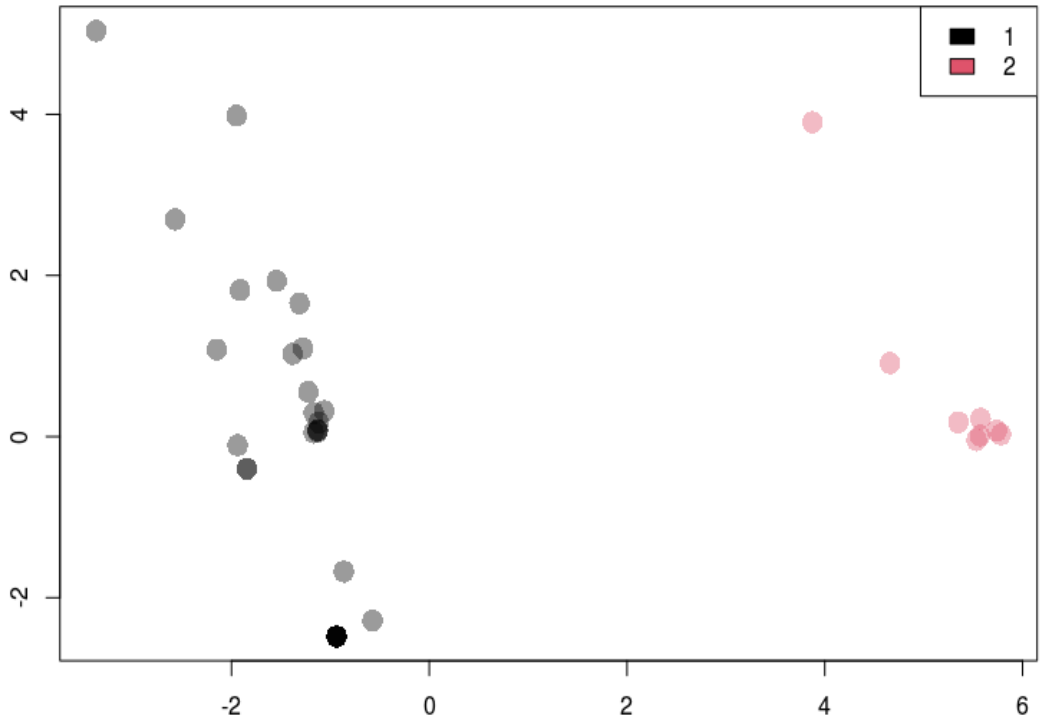
Os comandos utilizados no R para agrupar as bactérias de acordo com a presença ou ausência e quantidade de inserção de fago nas bactérias pode ser acessado em https://github.com/anelizebaranzeli1997/Comandos-pamk.

FIGURA 8 - Micobactérias agrupadas pela quantidade de inserção de micobacteriófagos específicos utilizando o método PCA. A cor preta corresponde ao grupo 1 do pamk e às micobactérias **M. marinum**, **M. sp. MCS**, **M. ulcerans**, **M. sp. JLS**, **M. avium**, **M. avium ssp. paratuberculosis K-10**, **M. sp. KMS**, **M. tuberculosis H37Ra**, **M. africanum**, **M. tuberculosis CDC1551**, **M. bovis BCG**, **M. tuberculosis F11** e **M. vanbaalenii**. A cor vermelha corresponde ao grupo 2 do pamk e às micobactérias **M. avium ssp. Paratuberculosis MAP/TANUVAS/TN/India/2008**, **M. liflandii**, **M. lepraemurium**, **M. fortuitum**, **M. tuberculosis H37Rv**, **M. microti**, **M. tuberculosis Haarlem**, **M. bovis**, **M. phlei**, **M. shigaense**, **M. tuberculosis C**, **M. vaccae**, **M. gilvum**, **M. stephanolepidis**, **M. abscessus** e **M. intracellulare**. A cor verde corresponde ao grupo 3 do pamk e às micobactérias **M. simiae**, **M. fallax**, **M. thermoresistibile**, **M. flavescens**, **M. chelonae**, **M. leprae**, **M. senegalense** e **M. neoaurum**. Por fim, a cor azul corresponde ao grupo 4 do pamk e às micobactérias **M. dioxanotrophicus** e **M. smegmatis**.



Fonte: Elaborada pela autora, 2021.

FIGURA 9 - Micobactérias agrupadas pela presença ou ausência de inserção de micobacteriófagos utilizando o método PCA. A cor preta corresponde ao grupo 1 do pamk e às micobactérias **M. marinum**, **M. sp. MCS**, **M. ulcerans**, **M. sp. JLS**, **M. avium**, **M. avium ssp. Paratuberculosis K-10**, **M. sp. KMS**, **M. tuberculosis H37Ra**, **M. africanum**, **M. tuberculosis CDC1551**, **M. bovis bcg**, **M. tuberculosis F11**, **M. vanbaalenii**, **M. avium ssp. Paratuberculosis MAP/TANUVAS/TN/India/2008**, **M. liflandii**, **M. lepraemurium**, **M. fortuitum**, **M. tuberculosis H37Rv**, **M. microti**, **M. tuberculosis Haarlem**, **M. bovis**, **M. phlei**, **M. shigaense**, **M. tuberculosis C**, **M. vaccae**, **M. gilvum**, **M. stephanolepidis**, **M. abscessus**, **M. intracellulare**, **M. dioxanotrophicus** e **M. smegmatis**. A cor vermelha corresponde ao grupo 2 do pamk e às micobactérias **M. senegalense**, **M. leprae**, **M. chelonae**, **M. flavescens**, **M. simiae**, **M. thermoresistibile**, **M. neoaurum** e **M. fallax**.



Fonte: Elaborada pela autora, 2021.

Observando que a quantidade de grupos entre as micobactérias agrupadas pela quantidade de inserção de micobacteriófagos específicos e as micobactérias agrupadas pela presença ou ausência de inserção de micobacteriófago difere, sugere-se que o

agrupamento das micobactérias pela quantidade não se dá pela presença de micobacteriófago específico dentro dos grupos.

4.8 Análise de enriquecimento para vias metabólicas dos insertos posicionados em proteínas de micobactérias por grupo.

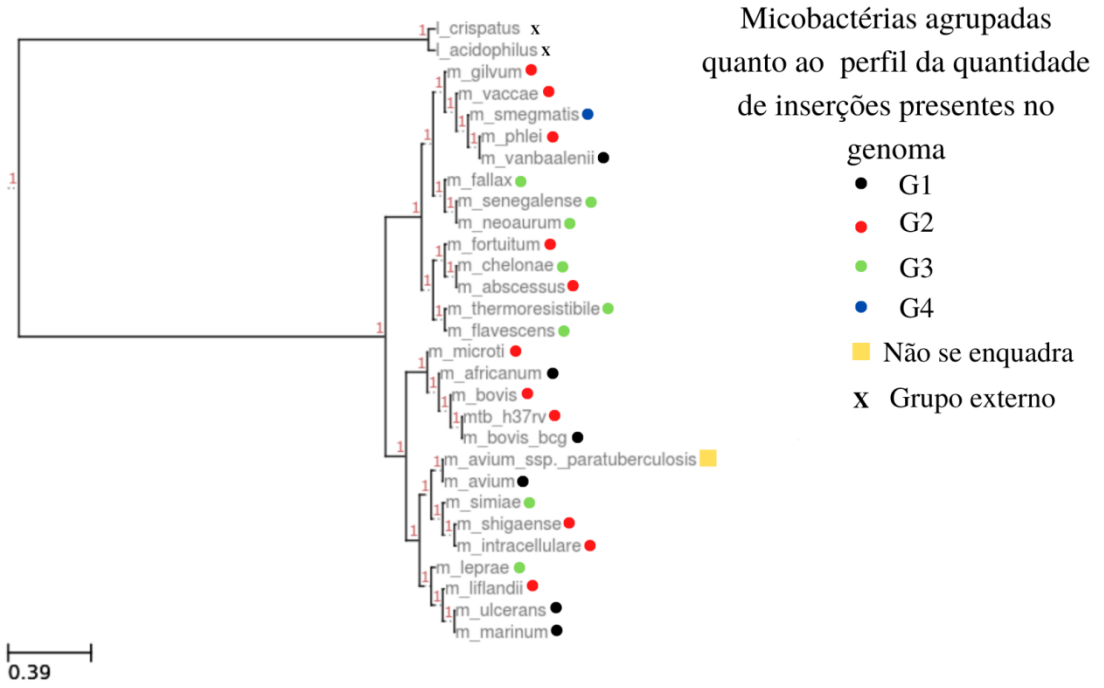
As proteínas obtidas no tópico 4.4 foram divididas de acordo com a bactéria de origem da inserção e a que grupo de bactérias agrupadas pela quantidade de inserção de fagos específicos elas pertencem, isso pode ser visualizado na tabela a seguir https://github.com/anelizebaranzeli1997/Proteinas-divididas-por-grupo**.**

A partir dessa divisão, para verificar se dentro dos grupos existe algum termo que apareça mais dentro de um grupo do que de outros, foram feitas análises de enriquecimento por grupo utilizando o ShinyGO, não tendo enriquecimento para vias metabólicas, como GO e Pfam.

4.9 Cladograma das micobactérias

Para o estudo da relação filogenética entre os grupos por quantidade de inserção de micobacteriófagos foi feito, baseando-se no gene rpoB de micobactérias, o cladograma abaixo.

FIGURA 10 - Cladograma das micobactérias. Os grupos fazem referência a análise feita utilizando o método pamk, que separou as micobactérias em grupos de acordo com a quantidade de inserções presentes no genoma. A cor preta corresponde ao grupo 1, a cor vermelha ao grupo 2, a cor verde ao grupo 3 e a cor azul ao grupo 4. A sequência rpoB da micobactéria **M. avium ssp. paratuberculosis** não especifica a qual cepa pertence, como no trabalho foram utilizadas duas diferentes, não foi possível enquadrá-la segundo os grupos analisados.



Fonte: Elaborada pela autora, 2021.

A seleção do grupo externo foi feita procurando-se por rpoB em todos os organismos na busca avançada do NCBI (COORDINATORS, 2016).

Os arquivos de saída obtidos com o alinhamento Clustal Omega, que realiza um alinhamento global, ou seja, alinhamento de toda a extensão da sequência (SIEVERS; HIGGINS, 2014), podem ser conferidos em https://github.com/anelizebaranzeli1997/Arquivos-Clustal.

A estatística foi realizada utilizando o pacote pvclust (SUZUKI et al., 2006) do R (R CORE TEAM, 2014), que utiliza o método de clusterização hierárquica, onde os objetos de estudo, nesse caso as micobactérias, semelhantes são agrupados em clusters, esse processo se dá com a identificação dos clusters mais próximos e a união desses clusters até que todos os clusters estejam mesclados (MURTAGH; CONTRERAS,

2012). As linhas de comando utilizadas podem ser verificadas em https://github.com/anelizebaranzeli1997/Pvclust.

Os resultados obtidos foram satisfatórios, já que a medida de suporte para cada nó foi igual a 1, o valor máximo possível, significando alta probabilidade de que os organismos que partem de cada nó se agrupam excluindo outras possibilidades. O valor de 0.39 representa uma escala para a quantidade de mudanças genéticas (RAMBAUT, 2015). O cladograma foi capaz de dividir as micobactérias entre as de crescimento rápido e lento, além de concentrar em um ramo as micobactérias do Complexo **Mycobacterium tuberculosis** (BROSCH et al., 2002, FUKANO et al., 2018, KHAN et al., 2002, KIM et al., 1999, LAVANIA et al., 2014, MVE-OBIANG et al., 2005).

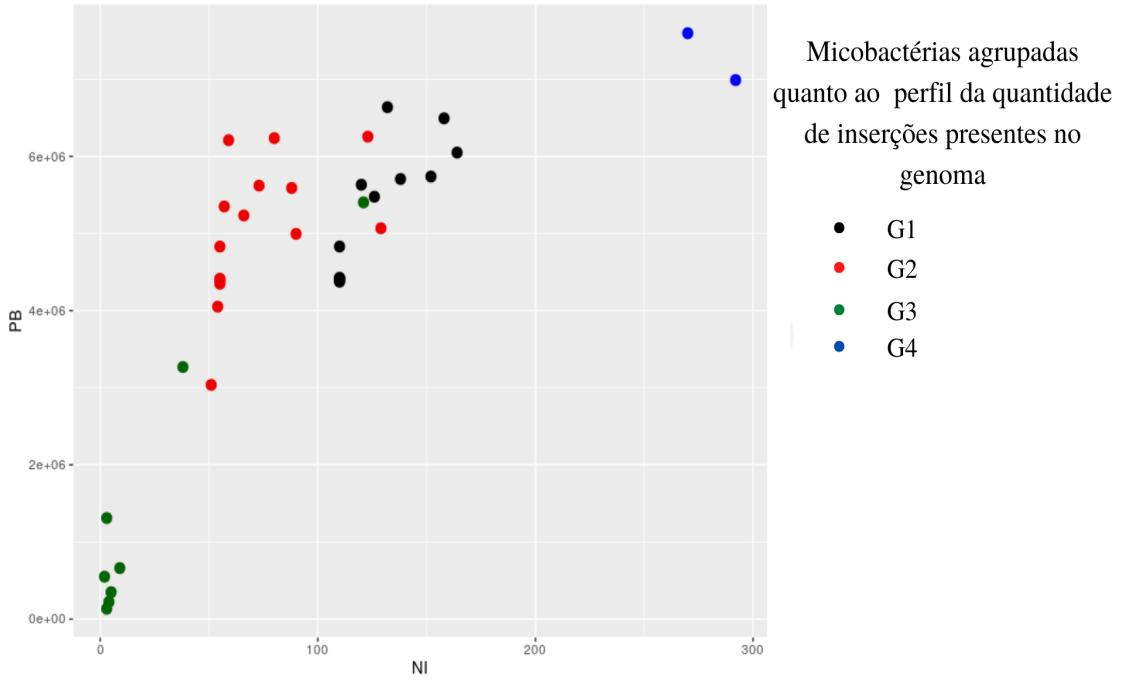
4.10 Avaliação da relação entre tamanho do genoma e quantidade de inserções por grupo de micobactérias agrupadas pela quantidade de inserção de micobacteriófagos específicos.

Observando-se que um grupo é formado pelas micobactérias de maior tamanho de genoma, a hipótese de que há relação entre tamanho do genoma e quantidade de inserções foi levantada.

Para realizar esse estudo criou-se uma tabela com as informações de micobactéria, o tamanho do seu genoma e a qual grupo ela pertence disponível em https://github.com/anelizebaranzeli1997/Bacteria-Tamanho-do-Genoma-Grupo.

Utilizando as linhas de comando contidas no documento disponível em https://github.com/anelizebaranzeli1997/Comandos-Grafico-de-Dispersao, obteve-se o gráfico a seguir que apresenta visível linearidade entre as variáveis indicando uma possível correlação entre elas.

FIGURA 11 - Gráfico de dispersão entre tamanho do genoma e quantidade de inserções de fago específico por grupo. PB é referente a quantidade de pares de base e NI ao número de inserções. Os grupos fazem referência a análise feita utilizando o método pamk, que separou as micobactérias em grupos de acordo com a quantidade de inserções presentes no genoma, sendo o grupo 1 do pamk representado aqui pela cor preta, o grupo 2 pela cor vermelha, o 3 pela cor verde e o 4 pela cor azul.



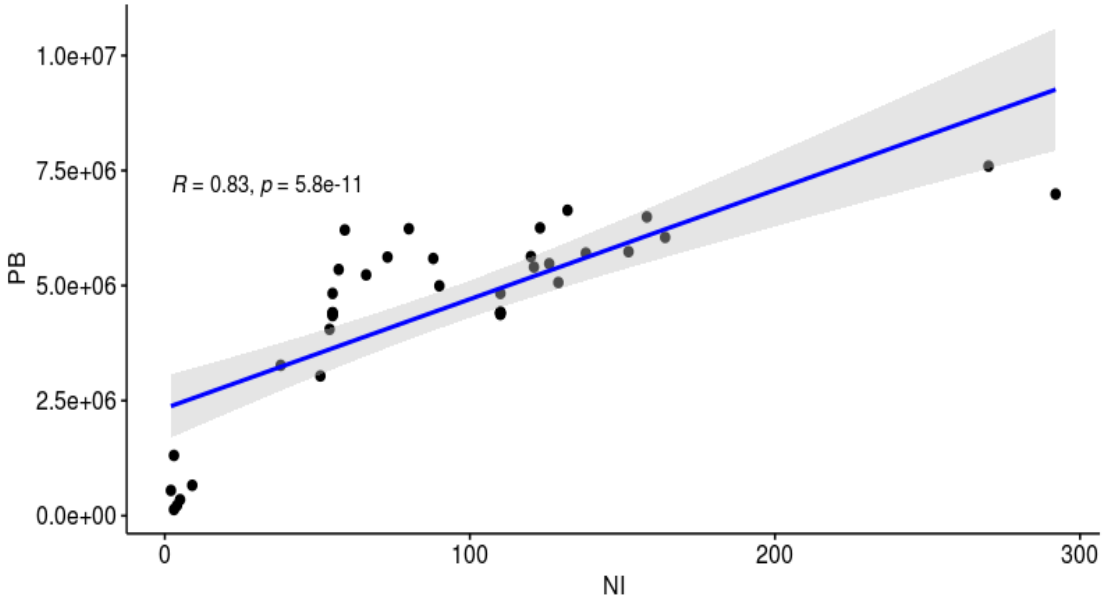
Fonte: elaborada pela autora, 2021.

4.11 Teste de normalidade e de correlação

Com os dados de tamanho do genoma das micobactérias e quantidade de inserções de micobacteriófago específico por micobactéria foi realizado o teste de Shapiro-Wilk (SHAPIRO; WILK, 1965) para testar a relação entre as duas variáveis, obtendo-se para os dados de tamanho do genoma um p-value = 0.00299 e para os dados de quantidade de inserções de fago específico por micobactéria um p-value = 0.002207.

O resultado do coeficiente de correlação de Spearman (SPEARMAN, 1961) pode ser observado no gráfico abaixo. As linhas de comando podem ser conferidas em https://github.com/anelizebaranzeli1997/Comandos-Teste-de-Normalidade-e-Spearman.

FIGURA 12 - Gráfico da correlação de spearman. PB é referente a quantidade de pares de base e NI ao número de inserções. Os círculos representam a distribuição das micobactérias no gráfico.



Fonte: Elaborado pela autora, 2021.

5. DISCUSSÃO

Bactérias são sequenciadas rapidamente, e é comum que se encontre genomas de fagos integrados ao genoma dessas bactérias, sendo eles capazes de participar de importantes propriedades biológicas, o que torna essencial o estudo dessas inserções para entender os genomas bacterianos sequenciados, incluindo compreender a relação evolutiva entre os dois organismos (CASJENS, 2003). Neste trabalho, utilizando-se 39 sequências de genomas completos de micobactérias e 512 genomas completos de micobacteriófagos, foram identificadas 4.593 regiões de genoma de micobacteriófago no genoma de micobactérias, sendo identificados ao menos um alinhamento em cada genoma de micobactéria utilizada no trabalho.

Observando o tamanho do alinhamento podemos identificar que existem muitas inserções com tamanho de alinhamento no intervalo que o CRISPRFinder considera como tamanho de espaçadores, entre 25 e 60 pares de base (GRISSA et al., 2007), além do fato de que o sistema CRISPR é uma forma de estudar a coevolução, já que as bactérias incorporam o material genético de bacteriófagos como espaçadores, sendo essa incorporação de sequência uma forma de tornar essas bactérias resistentes aos bacteriófagos que possuem essa sequência incorporada (VALE; LITTLE, 2010).

Para esse estudo, foi necessário o desenvolvimento do segundo programa do trabalho, derivado do primeiro, o getPosition100.py, que adiciona 100 pares de base a montante e a jusante do inserto. Dentro do conjunto de dados obtidos através do CRISPRFinder e os insertos de micobacteriófagos em micobactérias não foi encontrado nenhum alinhamento, dessa maneira não foi possível estabelecer uma relação entre os insertos e CRISPR de micobactérias.

Com um total de 64 sequências de proteínas de micobactérias e 4.593 insertos provenientes de micobacteriófagos no genoma de micobactérias obteve-se um total de 852 alinhamentos. Sendo 53 das 64 sequências de proteínas alinhadas a algum inserto, e destes, 534 possuem 100% de alinhamento. Como a relação entre proteínas dos

bacteriófagos e as proteínas das bactérias é importante para que haja infecção (ROUCOURT; LAVIGNE, 2009) era esperado que houvesse insertos de micobacteriófagos nas proteínas de micobactérias. Também se nota uma grande quantidade de proteínas alinhadas aos insertos sendo relacionadas aos micobcateriófagos, como por exemplo proteína da cauda menor do fago, uma proteína envolvida com a montagem da cauda em bacteriófagos, tendo a cauda um papel fundamental na infecção por fagos, pois participa do processo de penetração da parede celular bacteriana (HOFER, 2016).

As 53 sequências de proteínas que tiveram algum alinhamento com insertos foram submetidas a uma análise de enriquecimento para termos do GO utilizando a plataforma Blast2go (CONESA et al., 2005). Dessa análise obteve-se dois resultados, uma proteína está envolvida em processos biológicos, onde o gene contribui para um fim biológico, com processos comumente envolvendo transformações físicas ou químicas, onde o que entra no processo sai diferente após as mudanças, e a outra com função molecular, relacionada a atividade bioquímica de um produto gênico (ASHBURNER, 2000).

A presença ou ausência de micobacteriófago específico e quantidade de inserção de micobacteriófago nas bactérias permitiu classificar as micobactérias em quatro grupos segundo a quantidade de inserção e em dois grupos segundo a presença ou ausência de micobacteriófago específico, esse último, apresenta o grupo 2 sendo composto pelas micobactérias do grupo 3 na análise segundo a quantidade de inserção de micobacteriófagos nas bactérias, sendo esse grupo 3 composto, em sua maioria, pelas micobactérias com menores genomas e menores quantidade de inserções, o que sugere que não seria possível dividir as micobactérias em grupos segundo a presença ou ausência de micobacteriófago específico.

A análise de enriquecimento é capaz de identificar dentro do grupo de genes se há termos que estão mais presentes nesse grupo do que estariam ao acaso (TIPNEY; HUNTER, 2010). Caso houvesse enriquecimento de termos para as proteínas presentes

em cada grupo de micobactéria, seria possível observar a preferência dos micobacteriófagos por proteínas específicas dentro de cada grupo, porém, a hipótese foi rejeitada, já que não houve o enriquecimento de termos.

Como os micobacteriófagos evoluem de maneira muito rápida, as bactérias desenvolveram mecanismos para tentar escapar da infecção, o que leva a coevolução (STERN, 2011), então o agrupamento das micobactérias em quatro grupos diferentes foi estudado de maneira a entender se interferiria na filogenia das micobactérias. O cladograma foi construído baseando-se no gene rpoB de micobactérias, que é uma região bem conservada entre as eubactérias e já utilizada para filogenia em micobactérias (KIM et al., 1999, ADÉKAMBI, 2003). Não foi possível observar relação entre as bactérias que constituem cada um dos quatro grupos formados pelas bactérias agrupadas pela quantidade de inserção de micobacteriófagos específicos com os nós do cladograma.

Foi observado nos agrupamentos por inserção, o grupo 4 é formado por somente duas micobactérias, a **Mycobacterium dioxanotrophicus** e a **Mycobacterium smegmatis**, sendo as duas micobactérias com o maior tamanho de genoma dentro das estudadas no presente trabalho, com 7.595.921 e 6.988.209 pares de base respectivamente, e para testar a hipótese de correlação entre essas duas variáveis, primeiro um gráfico de dispersão foi feito baseando-se no tamanho do genoma e quantidade de inserções, dividindo as micobactérias pelos grupos a qual elas pertencem na figura 4, onde é possível observar uma linearidade positiva no gráfico, então se fez necessário testar estatisticamente a correlação entre as duas variáveis.

O resultado da correlação de Spearman indica uma forte correlação entre o tamanho do genoma e a quantidade de inserções (SPEARMAN, 1961), sugerindo que quanto maior o tamanho do genoma, maior é a quantidade de inserções totais, independente de micobacteriófago específico, que a micobactéria terá.

6 CONCLUSÃO

Foi possível levantar o genoma completo de 39 micobactérias e 512 micobacteriófagos, havendo inserção de micobacteriófagos em todas as sequências de micobactérias utilizadas no presente trabalho que variam em tamanho entre 28 a 5040 pb.

O desenvolvimento dos programas getPosition.py e getPosition100.py permitiram a localização dessas inserções dentro do genoma das micobactérias.

As inserções foram analisadas, *in silico*, utilizando o CRISPRFinder, sugerindo que essas regiões não são componentes de um sistema CRISPR de micobactérias. Algumas fazem parte de proteínas de micobactérias, mas não possuem enriquecimento para termos do GO (Blast2go).

Foi possível classificar as micobactérias em grupos segundo a quantidade de inserção de micobacteriófagos.

A presença ou ausência de micobacteriófago específico não foi o motivo da separação das micobactérias. Não existe enriquecimento para vias metabólicas dentro dos grupos. A filogenia não apresenta táxons em comum com a divisão de quatro grupos das micobactérias. Encontrou-se uma forte relação entre o tamanho do genoma e a quantidade de inserções totais observadas no gráfico de dispersão que foi confirmada com a correlação de Spearman.

Com o presente trabalho foi possível estudar a coevolução entre os dois organismos de maneira totalmente *in silico*, com a criação de ferramentas que são capazes auxiliar trabalhos futuros com outros conjuntos de dados. Podendo estabelecer outras relações coevolutivas utilizando-se deles para o estudo das regiões específicas de inserção dentro do genoma de um organismo, visto que são capazes de separar esse pedaço do genoma do restante para as análises mais detalhadas que forem necessárias.

7 REFERÊNCIAS

ADÉKAMBI, Toïdi; COLSON, Philippe; DRANCOURT, Michel. rpoB-based identification of nonpigmented and late-pigmenting rapidly growing mycobacteria. **Journal of clinical microbiology**, v. 41, n. 12, p. 5699-5708, 2003.

ALTSCHUP, S. F. et al. Altschul-1990-Basic Local Alignment. p. 403–410, 1990.

ASHBURNER, Michael et al. Gene ontology: tool for the unification of biology. **Nature genetics**, v. 25, n. 1, p. 25-29, 2000.

BATEMAN, A. UniProt: A worldwide hub of protein knowledge. **Nucleic Acids Research**, v. 47, n. D1, p. D506–D515, 2019.

BOYD, E. F.; DAVIS, B. M.; HOCHHUT, B. Bacteriophage-bacteriophage interactions in the evolution of pathogenic bacteria. **Trends in Microbiology**, v. 9, n. 3, p. 137–144, 2001.

BRASIL. Ministério da Saúde. Agência Nacional de Vigilância Sanitária. Manual de Bacteriologia da tuberculose. 3ª ed. Rio de Janeiro, 2005. Disponível em: http://www.saude.mt.gov.br/upload/documento/81/manual-de-bacteriologia-da-tuberculose- %5B81-080909-SES-MT%5D.pdf. Acesso em: set/2019.

BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Vigilância Epidemiológica. Manual Nacional de Vigilância Laboratorial da Tuberculose e outras Micobactérias. 1ª ed. Brasília, 2008. Disponível em: http://bvsms.saude.gov.br/bvs/publicacoes/manual_vigilancia_laboratorial_tuberculose.pdf. Acesso em set/2019.

BRASIL. Agência Nacional de Vigilância Sanitária. Nota técnica conjunta Nº 01/2009. Infecções por micobactérias de crescimento rápido: fluxo de notificações, diagnósticos clínico, microbiológico e tratamento, 2009. Disponível em:

http://www.saude.sp.gov.br/resources/cve-centro-de-vigilancia-epidemiologica/areas-de vigilancia/infeccao-hospitalar/doc/nt0109_conjunta.pdf. Acesso em: set/2019.

BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Vigilância das Doenças Transmissíveis.Brasil Livre da Tuberculose : Plano Nacional pelo Fim da Tuberculose como Problema de Saúde Pública. Brasília, 2017. Disponível em: https://portalarquivos2.saude.gov.br/images/pdf/2017/julho/05/af-miolo-plano-nac tuberculose-29jun17-grafica.pdf. Acesso em: set/2019.

BROSCH, R. et al. A new evolutionary scenario for the Mycobacterium tuberculosis complex. **Proceedings of the National Academy of Sciences, USA**, v. 99, n. 6, p. 3684–3689, 2002.

CASJENS, Sherwood. Prophages and bacterial genomics: what have we learned so far?. **Molecular microbiology**, v. 49, n. 2, p. 277-300, 2003.

COLLINS, C. H.; GRANGE, J. M.; YATES, M. D. Mycobacteria in water. **Journal of Applied Bacteriology**, v. 57, n. 2, p. 193–211, 1984.

COMAS, Ĩ. et al. Human T cell epitopes of Mycobacterium tuberculosis are evolutionarily hyperconserved. **Nature Genetics**, v. 42, n. 6, p. 498–503, 2010.

CONESA, A. et al. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. **Bioinformatics**, v. 21, n. 18, p. 3674–3676, 2005.

COORDINATORS, N. R. Database resources of the National Center for Biotechnology Information [ftp://ftp.ncbi.nih.gov/genomes/Bacteria]. **Nucleic Acids Research**, v. 44, n. D1, p. D7–D19, 2016.

DE FREITAS ALMEIDA, Gabriel Magno; SUNDBERG, Lotta-Riina. The forgotten tale of Brazilian phage therapy. **The Lancet Infectious Diseases**, v. 20, n. 5, p. e90-e101, 2020.

DESIERE, F. et al. Comparative genomics reveals close genetic relationships between phages from dairy bacteria and pathogenic streptococci: Evolutionary implications for prophage-host interactions. **Virology**, v. 288, n. 2, p. 325–341, 2001.

DIAS DE ARAÚJO, N. et al. A era da bioinformática: seu potencial e suas implicações para as ciências da saúde. **Estudos de biologia**, v. 30, n. 70/72, p. 143-148, 2008.

DUNTEMAN, George H. **Principal components analysis**. Sage, 1989.

FREY, C. A. A METHOD FOR DETECTING ACID-FAST BACTERIA IN THE SOIL. **Science**, v. 71, n. 1840, p. 366–366, 1930.

FUKANO, Hanako et al. Mycobacterium shigaense sp. nov., a slow-growing, scotochromogenic species, is a member of the Mycobacterium simiae complex. **International journal of systematic and evolutionary microbiology**, v. 68, n. 8, p. 2437-2442, 2018.

GE, S. X. et al. ShinyGO: A graphical gene-set enrichment tool for animals and plants. **Bioinformatics**, v. 36, n. 8, p. 2628–2629, 2020.

GENTILE, Gabrielle M. et al. More evidence of collusion: a new prophage-mediated viral defense system encoded by mycobacteriophage Sbash. **MBio**, v. 10, n. 2, p. e00196-19, 2019.

GRISSA, I.; VERGNAUD, G.; POURCEL, C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. **Nucleic Acids Research**, v. 35, n. Web Server issue, p. 52–57, 2007.

HATFULL, G. F. et al. Comparative Genomic Analysis of 60 Mycobacteriophage Genomes: Genome Clustering, Gene Acquisition, and Gene Size. **Journal of Molecular Biology**, v. 397, n. 1, p. 119–143, 2010.

HATFULL, Graham F. Mycobacteriophages. **Microbiology spectrum**, v. 6, n. 5, p. 6.5. 08, 2018.

HENNIG, Christian; IMPORTS, M. A. S. S. Package ‘fpc’. **Available at: https://cran. r-project. org/web/packages/fpc/index. html ENT**, v. 91, 2015.

HERSHBERG, R. et al. High functional diversity in Mycobacterium tuberculosis driven by genetic drift and human demography. **PLoS Biology**, v. 6, n. 12, p. 2658–2671, 2008.

HOFER, Ursula. The sting is in the phage's tail. **Nature Reviews Microbiology**, v. 14, n. 8, p. 477-477, 2016.

HUERTA-CEPAS, J.; SERRA, F.; BORK, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. **Molecular Biology and Evolution**, v. 33, n. 6, p. 1635–1638, 2016.

ISHINO, Y. et al. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isoenzyme conversion in Escherichia coli, and identification of the gene product. **Journal of Bacteriology**, v. 169, n. 12, p. 5429–5433, 1987.

JANSEN, R. et al. Identification of genes that are associated with DNA repeats in prokaryotes. **Molecular Microbiology**, v. 43, n. 6, p. 1565–1575, 2002.

KATOCH, V. M. Infections due to non-tuberculous mycobacteria (NTM). **Indian Journal of Medical Research**, v. 120, p. 290-304, 2004.

KAUFMAN, Leonard; ROUSSEEUW, Peter J. **Finding Groups in Data: An Introduction to Cluster Analysis**. [*S. l.*: *s. n.*], 1990.

KHAN, Ashraf A. et al. Classification of a polycyclic aromatic hydrocarbon-metabolizing bacterium, Mycobacterium sp. strain PYR-1, as Mycobacterium vanbaalenii sp. nov. **International Journal of Systematic and Evolutionary Microbiology**, v. 52, n. 6, p. 1997-2002, 2002.

KIM, B. J. et al. Identification of mycobacterial species by comparative sequence analysis of the RNA polymerase gene (rpoB). **Journal of clinical microbiology**, v. 37, n. 6, p. 1714–1720, 1999.

MORGADO, Sergio Mascarenhas et al. Diversidade, taxonomia genômica e resistoma de micobactérias da Mata Atlântica. Tese de Doutorado. **INSTITUTO OSWALDO CRUZ**, 2017.

LAVANIA, Mallika et al. Detection of Mycobacterium gilvum first time from the bathing water of leprosy patient from Purulia, West Bengal. **International journal of mycobacteriology**, v. 3, n. 4, p. 286-289, 2014.

MURTAGH, F.; CONTRERAS, P. Algorithms for hierarchical clustering: An overview. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 2, n. 1, p. 86–97, 2012.

MVE-OBIANG, Armand et al. A newly discovered mycobacterial pathogen isolated from laboratory colonies of Xenopus species with lethal infections produces a novel form of mycolactone, the Mycobacterium ulcerans macrolide toxin. **Infection and immunity**, v. 73, n. 6, p. 3307-3312, 2005.

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION (US); CAMACHO, Christiam. **BLAST (r) Command Line Applications User Manual**. National Center for Biotechnology Information (US), 2008.

PALOMINO, Juan Carlos *et al*. **From basic science to patient care**. 1. ed. [*S. l.*: *s. n.*], 2007.

PANDOLFI, J. R. et al. Tuberculose e o estudo molecular da sua epidemiologia. **Revista de Ciências Farmacêuticas Básica e Aplicada**, v. 28, n. 3, p. 251–257, 2007.

PEARSON, K. LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, v. 2, n. 11, p. 559–572, 1901.

PEDULLA, M. L. et al. Origins of Highly Mosaic Mycobacteriophage Genomes. **Cell,** v. 113, n. 2, p. 171–182, 2003.

PETERS, D. L. et al. Novel Stenotrophomonas maltophilia temperate phage DLP4 is capable of lysogenic conversion. **BMC Genomics**, v. 20, n. 1, p. 1–17, 2019.

POPE, W. H. et al. Whole genome comparison of a large collection of mycobacteriophages reveals a continuum of phage genetic diversity. **eLife**, v. 4, n. 4, p. 1-65, 2015.

PÔSSA, Tâmea Aparecida Linhares. Infecções de pele e partes moles causadas por Mycobacterium Abscessus após procedimentos cirúrgicos estéticos: análise de aspectos clínicos, terapêuticos e microbiológicos. Dissertação de Mestrado. **Universidade Federal do Espírito Santo**, 2011.

R Core Team. R: A Language and Environment for Statistical Computing. **R Foundation for Statistical Computing.** Vienna, Austria, 2014.

RAMBAUT, Andrew. **How to read a phylogenetic tree**. [*S. l.*], 30 jul. 2015. Disponível em: https://artic.network/how-to-read-a-tree.html. Acesso em: 3 ago. 2021.

RODRIGUES, Fabiene Silva. **Métodos de agrupamento na análise de dados de expressão gênica**. 2009. Dissertação (Mestrado em Estatística) - Pós-Graduação em Estatística, Departamento de Estatística, Universidade Federal de São Carlos, São Carlos - São Paulo, 2009.

ROUCOURT, B.; LAVIGNE, R. The role of interactions between phage and bacterial proteins within the infected cell: A diverse and puzzling interactome. **Environmental Microbiology**, v. 11, n. 11, p. 2789–2805, 2009.

SANTANA, JÚLIA CARDOSO; SILVA, CLAÚDIA PERES DA; PEREIRA, C. A. Principais doenças oportunistas em indivíduos com HIV. **Humanidades & Tecnologia em Revista (FINOM)**, v. 1, n. 16, p. 405–422, 2019.

SANTOS, Fabrício R.; ORTEGA, José Miguel. Bioinformática aplicada à Genômica. **Melhoramento Genômico, Minas Gerais: UFV**, p. 93-98, 2003.

SANTOS, Mariana Oliveira et al. **Micobactérias: identificação e perfil de sensibilidade a tuberculostáticos em amostras isoladas no Laboratório Central de Saúde Pública do Estado do Piauí, janeiro 2014 a março de 2015**. 2015. Tese de Doutorado.

SASSI, M. et al. Mycobacteriophage-drived diversification of Mycobacterium abscessus. **Biology Direct**, v. 9, n. 1, 2014.

SHAPIRO, S. S.; WILK, M. B. An Analysis of Variance Test for Normality (Complete Samples). **Biometrika**, v. 52, n. 3/4, p. 591, 1965.

SIEVERS, F.; HIGGINS, D. G. Clustal Omega. **Current Protocols in Bioinformatics**, v. 2014, n. December, p. 3.13.1-3.13.16, 2014.

SPEARMAN, Charles. The proof and measurement of association between two things. 1961.

STERN, Adi; SOREK, Rotem. The phage‐host arms race: shaping the evolution of microbes. **Bioessays**, v. 33, n. 1, p. 43-51, 2011.

SUZUKI, R.; SHIMODAIRA, H. Pvclust: An R package for assessing the uncertainty in hierarchical clustering. **Bioinformatics**, v. 22, n. 12, p. 1540–1542, 2006.

THOMPSON, J. N. The coevolutionary process. **The University of Chicago Press Chicago and London**, 1994.

TIPNEY, Hannah; HUNTER, Lawrence. An introduction to effective use of enrichment analysis software. **Human genomics**, v. 4, n. 3, p. 1-5, 2010.

UEKI, Suely Yoko Mizuka et al. Micobactérias não-tuberculosas: diversidade das espécies no estado de São Paulo. **J Bras Patol Med Lab**, v. 41, n. 1, p. 1-8, 2005.

VALE, P. F.; LITTLE, T. J. CRISPR-mediated phage resistance and the ghost of coevolution past. **Proceedings of the Royal Society B: Biological Sciences**, v. 277, n. 1691, p. 2097–2103, 2010.

VALERO-MORA, P. M. ggplot2: Elegant Graphics for Data Analysis . **Journal of Statistical Software**, v. 35, n. Book Review 1, 2010.

VAN ROSSUM, G. Python tutorial, Technical Report CS-R9526. **Centrum voor Wiskunde en Informatica (CWI)**, 1995.

VAN VALEN, Leigh. A new evolutionary law. **Evol Theory**, v. 1, p. 1-30, 1973.

WALDOR, M. K.; MEKALANOS, J. J. Lysogenic Conversion by a Filamentous Phage Encoding Cholera Toxin. **Science**, v. 272, n. 5270, p. 1910–1914, 1996.

WIRTH, T. et al. Origin, spread and demography of the Mycobacterium tuberculosis complex. **PLoS Pathogens**, v. 4, n. 9, 2008.