UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ARTES, CIÊNCIAS E HUMANIDADES
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

PRISCILLA KOCH WAGNER

**Uma nova abordagem para identificação da provável origem de genes
exclusivos de bactérias**

São Paulo

2018

PRISCILLA KOCH WAGNER

**Uma nova abordagem para identificação da provável origem de genes
exclusivos de bactérias**

Versão corrigida

Dissertação apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação.

Área de concentração: Metodologia e Técnicas da Computação

Versão corrigida contendo as alterações solicitadas pela comissão julgadora em 26 de março de 2018. A versão original encontra-se em acervo reservado na Biblioteca da EACH-USP e na Biblioteca Digital de Teses e Dissertações da USP (BDTD), de acordo com a Resolução CoPGr 6018, de 13 de outubro de 2011.

Orientador: Prof. Dr. Luciano Antonio Digiampietri

São Paulo

2018

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

CATALOGAÇÃO-NA-PUBLICAÇÃO

(Universidade de São Paulo. Escola de Artes, Ciências e Humanidades. Biblioteca)
CRB-8 4936

Wagner, Priscilla Koch

Uma nova abordagem para identificação da provável origem de genes exclusivos de bactérias / Priscilla Koch Wagner ; orientador Luciano Antonio Digiampietri. – 2018.

104 f. : il.

Dissertação (Mestrado em Ciências) - Programa de Pós-Graduação em Sistemas de Informação, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo.
Versão corrigida

1. Bioinformática. 2. Genomas - Simulação computacional. 3. Genomas - Comparação. 4. Filogenia. 5. Xanthomonas. I. Digiampietri, Luciano Antonio, orient. II. Título.

CDD 22.ed.– 572.80285

Dissertação de autoria de Priscilla Koch Wagner, sob o título “**Uma nova abordagem para identificação da provável origem de genes exclusivos de bactérias**”, apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação, na área de concentração Metodologia e Técnicas da Computação, aprovada em 26 de março de 2018 pela comissão julgadora constituída pelos doutores:

Prof. Dr. Luciano Antonio Digiampietri

Escola de Artes, Ciências e Humanidades - Universidade de São Paulo

Presidente

Prof. Dr. Márcio Oikawa

Centro de Matemática, Computação e Cognição

Prof. Dr. André Fujita

Instituto de Matemática e Estatística - Universidade de São Paulo

Profª. Dra. Arianne Machado Lima

Escola de Artes, Ciências e Humanidades - Universidade de São Paulo

Dedico este trabalho aos meus pais, Claudete e Roque, e a minha irmã, Camilla, pelo apoio e carinho que sempre me deram.

Agradecimentos

Agradeço, em primeiro lugar, aos meus pais e minha irmã por todo amor, compreensão e incentivo que me deram desde sempre e que me possibilitaram chegar até este momento.

Agradeço especialmente ao meu orientador, Prof. Dr. Luciano Antonio Digiampietri, por toda a dedicação, apoio, conselhos e orientações dados durante esta fase de mestrado, por me apresentar a área de bioinformática e os desafios envolvidos.

Agradeço aos meus amigos, por entenderem e me apoiarem a concluir essa etapa importante. Em especial, às também mestrandas Bárbara, Jéssica e Vivian, pelos auxílios, angústias e incentivos compartilhados.

Aos demais alunos envolvidos e ao grupo de pesquisa, Giovani, Guilherme e Vivian, por me auxiliarem em dúvidas, conselhos e sugestões em várias etapas deste projeto.

Agradeço a todos os professores que me acompanharam nessa jornada, por contribuírem com a minha formação, especialmente aos professores do Programa de Pós-Graduação em Sistemas de Informação (PPgSI). Do curso de graduação em Sistemas de Informação, agradeço especialmente à Profa. Dra. Sarajane Marques Peres, por me acompanhar desde a iniciação científica, o que me fez criar gosto pela pesquisa e também pelo incentivo que me deu para ingressar no mestrado.

Por fim, agradeço ao meu namorado, André, pelo carinho, paciência e compreensão durante essa fase final de mestrado. Agradeço por me apoiar nos momentos difíceis que enfrentei nesses últimos meses, por ser meu suporte, meu melhor amigo e por não me deixar desistir em momento nenhum.

Resumo

WAGNER, Priscilla Koch. **Uma nova abordagem para identificação da provável origem de genes exclusivos de bactérias**. 2018. 104 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2018.

A comparação de genomas, genes ou até sequências de nucleotídeos não condificantes é uma importante tarefa na qual a bioinformática pode ser aplicada, uma vez que ela auxiliar em diversas atividades, por exemplo, análises filogenéticas. Análise filogenética, por sua vez, busca analisar a relação evolutiva de cada espécie, considerando suas características genéticas. Esses processos e as técnicas que os implementam se baseiam em sequências de nucleotídeos sequenciadas e armazenadas em bancos de dados de genomas públicos. Com análise filogenética também é possível identificar possíveis origens de um gene. Essa tarefa é de grande importância, pois auxilia na identificação da origem de genes patogênicos, podendo auxiliar no combate e prevenção do surgimento de doenças. Um problema potencial dessas sequências é a possibilidade de haver erros nas anotações (marcações de sequências como genes). Esses erros são pouco explorados por pesquisadores atualmente. Outro tema pouco explorado é a análise filogenética de genes exclusivos, que são genes que se manifestam em apenas uma espécie, considerando um grupo de espécies próximas. A identificação de genes exclusivos de alguma espécie pode servir para a correta identificação de, por exemplo, a espécie que causa uma doença, de forma a permitir o uso do tratamento mais específico e adequado. A importância da descoberta de filogenias de genes exclusivos e a dificuldade de garantir a consistência nas anotações genéticas motivaram este trabalho, que teve como objetivo implementar ferramentas para interpretar dados de comparação genética, identificando potenciais erros em anotação de genes exclusivos e criando estratégias para identificar a origem desses genes. As origens de genes exclusivos exploradas neste trabalho envolvem a possibilidade dos genes exclusivos terem derivado de outras famílias de genes do próprio organismo, ou, os genes exclusivos se diferenciaram muito dos genes ancestrais. Essas hipóteses, juntamente com a hipótese da existência de erros de anotação, foram exploradas em experimentos utilizando as ferramentas desenvolvidas. Os experimentos visaram a analisar a aplicabilidade da estratégia desenvolvida. Foram utilizados genomas de bactérias do gênero *Xanthomonas*, que contém um grande grupo de bactérias que causam doenças em plantas. Os resultados obtidos demonstram que existe uma quantidade considerável de potenciais erros de anotação nos genomas considerados, provando a hipótese de que a inconsistência nas anotações genômicas possui grande influência para a dificuldade na identificação de filogenias (tanto de genes exclusivos como para não exclusivos). Os resultados também demonstraram que boa parte dos genes exclusivos possivelmente se originaram de outras famílias de genes do próprio genoma. Ou ainda, que esses genes sofreram modificações em relação aos genes ancestrais, mas ainda possuem certas semelhanças com sequências de nucleotídeos que não codificam genes em outras espécies mais distantes. Por fim, a estratégia desenvolvida se mostrou útil na análise filogenética das bactérias estudadas, sendo este um forte indício de que a mesma abordagem pode ser utilizada para problemas similares com outras espécies de seres vivos.

Palavras-chaves: Bioinformática. Análise filogenética. Filogenia. Comparação de genes. Genes exclusivos. Anotação de genes. Genômica.

Abstract

WAGNER, Priscilla Koch. **A new approach to identify the probable origin of bacteria exclusive genes**. 2018. 104 p. Dissertation (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2018.

Comparison of genomes, genes or even non-coding nucleotide sequences is an important task in which bioinformatics can be applied, since it allows the application of phylogenetic analyses. Phylogenetic analysis, in its turn, seeks to analyze the evolutionary relation of each species, considering its genetic characteristics. These processes and the techniques that implement them are based on nucleotide sequences sequenced and stores in databases of public genomes. With phylogenetic analysis it is also possible to identify possible origins of a gene. This task has a great importance, because it allows the identification of the origin of pathogenic genes, which may help to combat or prevent diseases. A potential problem of these sequences is the possibility of having annotation errors (sequences marking as genes). These errors are little explored by researchers nowadays. Another unexplored topic is the phylogenetic analysis of exclusive genes, which are genes thaht manifest in only one species, considering a group of nearby species. The identification of exclusive genes of a species may serve to correctly identify, for example, a disease, in order to allow the use of a more especific and appropriate treatment. The importance of discovering phylogenies of exclusive genes and the difficulty of guaranteeing the consistency of genetic annotations motivated this work, whose objective was to implement tools to interpret data of genetic comparison, identifying annotation errors in exclusive genes and creating strategies to identify the origin of these genes. The origins of exclusive genes explored in this work involve the possibility of the exclusive genes have derived of other gene families of the organism itself, or, the exclusive genes differed a lot from the ancestral genes. Theses hypotheses, with the hypotesis of the existance of annotation errors, were explored in experiments using the developed tools. The experiments aimed to analyse the applicability of the developed strategy. Genomes of bacteria of the genus *Xanthomonas* were used, which contains a large group of bacteria that cause diseases in plants. The results show that there is a considerable amount of annotation errors on the genomes, proving the hypothesis that the inconsistency in genomic annotations has a great influence on the difficulty in identifying phylogenies (both exclusive and non-exclusive genes). The results also show that much of exclusive genes possibly originated from other gene families of the genome itself. Furthermore, these genes may have suffered modifications in relation to the ancestral genes, but still have certain similarities with nucleotide sequences that don't encode genes in other more distant species. Finally, the strategy developed proved useful on phylogenetic analysis of the studied bacteria, which is a strong indication that the same approach can be used for similar problems with other species of living beings.

Keywords: Bioinformatics. Phylogenetic analysis. Phylogeny. Gene comparison. Exclusive genes. Genes annotation. Genomic.

Lista de figuras

Figura 1 – Conceitos de genes homólogos, ortólogos e parálogos.	22
Figura 2 – Alinhamento entre quatro sequências de nucleotídeos. A parte <i>a</i> apresenta as sequências originais antes do alinhamento. A parte <i>b</i> apresenta o alinhamento das sequências, incluindo as lacunas adicionadas (hífens).	27
Figura 3 – Árvore filogenética de primatas: gibão preto (siamang), gibão (gibbon), orangotango (orangutan), gorila (gorilla), humano (human), chimpanzé (chimpanzee).	30
Figura 4 – Distribuição de artigos por categoria.	34
Figura 5 – Diagrama completo da abordagem desenvolvida nesse projeto.	46
Figura 6 – Exemplo da configuração de uma matriz de presença.	49
Figura 7 – Trechos de arquivos FASTA. (a) trecho no formato .FNA. (b) trecho no formato .FAA.	51
Figura 8 – Comparação de genomas e geração da matriz de presença com contagem de genes por genoma.	52
Figura 9 – Análise de inconsistências no processo de anotação genética.	54
Figura 10 – Análise de genes exclusivos alinhados com outras famílias de genes dentro do próprio genoma.	57
Figura 11 – Análise de genes exclusivos alinhados com sequências de nucleotídeos de outras espécies do banco do NCBI.	60
Figura 12 – Rede de genes das 15 espécies de <i>Xanthomonas</i> em análise neste estudo, apresentando as relações entre as diferentes famílias de genes homólogos dos 15 genomas.	65
Figura 13 – Filogenia de Referência	66
Figura 14 – Taxonomia da espécie <i>Xanthomonas campestris pv. campestris</i> , considerando 741 alinhamentos com o genoma da própria espécie.	67
Figura 15 – Distribuição de genes com erros intragenoma entre as 15 espécies de <i>Xanthomonas</i> consideradas.	70
Figura 16 – Espécies com erros intragenoma a mais que a quantidade de genes com este erro.	71

Figura 17 – Distribuição de genes envolvidos em potenciais erros intergenoma entre as 15 espécies de <i>Xanthomonas</i> consideradas.	72
Figura 18 – Relação entre os genes que apresentaram erro intergenoma e o total de genes considerados potencialmente exclusivos inicialmente.	73
Figura 19 – Distribuição de erros de anotação por quantidade de genomas em que o erro ocorreu.	74
Figura 20 – Frequência que cada genoma esteve envolvido em erros ou inconsistências de anotação em relação aos demais, considerando erros presentes em 3 genomas.	75
Figura 21 – Diagrama de <i>Chord</i> apresentando as inter-relações entre os genomas que possuem os genes anotados e os que apresentaram o erro.	76
Figura 22 – Diagrama de <i>Chord</i> apresentando os erros intergenoma entre a espécie <i>XccATCC33913</i> e os que apresentaram os erros.	77
Figura 23 – Genes cuja sequência de nucleotídeos com a qual alinhou possui códon de parada.	79
Figura 24 – Quantidade, por espécie, de genes exclusivos que não possuíam nenhum alinhamento com outra família de gene dentro do próprio genoma. . . .	81
Figura 25 – Relação entre os genes que não alinharam com outras famílias do próprio genoma e todos os genes exclusivos para cada espécie de <i>Xanthomonas</i>	81
Figura 26 – Quantidade, por espécie, de genes exclusivos que possuem alinhamento com outra família de gene dentro do próprio genoma.	82
Figura 27 – Relação entre os genes que possuem alinhamento com outras famílias do próprio genoma e todos os genes exclusivos para cada espécie de <i>Xanthomonas</i>	83
Figura 28 – Quantidade de alinhamentos dos genomas, pelos parâmetros de faixa de identidade e faixa de cobertura. Sendo: (a) <i>Xacitri</i> , (b) <i>Xacitrumelo</i> , (c) <i>Xalbilineans</i> , (d) <i>Xac29-1</i> , (e) <i>Xcc8004</i> , (f) <i>XccATC33913</i> , (g) <i>XccB100</i> e (h) <i>Xccitri</i>	84
Figura 29 – Quantidade de alinhamentos dos genomas, pelos parâmetros de faixa de identidade e faixa de cobertura. Sendo: (i) <i>Xraphani</i> , (j) <i>Xvesicatoria</i> , (k) <i>Xff</i> , (l) <i>XooKACC10331</i> , (m) <i>XooMAFF311018</i> , (n) <i>XooPXO99A</i> e (o) <i>Xooryzicola</i>	85
Figura 30 – Taxonomia menos abrangente da espécie <i>Xvesicatoria</i>	88

Figura 31 – Taxonomia mais abrangente da espécie <i>Xvesicatoria</i>	89
--	----

Lista de tabelas

Tabela 1 – Categorização dos artigos.	33
Tabela 2 – Genomas utilizados nos experimentos, com respectivas abreviações, seu tamanho e número de genes.	63
Tabela 3 – Relação entre a quantidade total de genes exclusivos e a quantidade total de genes de cada genoma	64
Tabela 4 – Comparação de genes exclusivos antes e após a análise, relação da quantidade de genes do genoma com os exclusivos encontrados e diminuição da quantidade de genes exclusivos.	77
Tabela 5 – Parâmetros e resultados das árvores menos abrangente para as 15 espécies de <i>Xanthomonas</i>	87
Tabela 6 – Parâmetros e resultados das árvores mais abrangente para as 15 espécies de <i>Xanthomonas</i>	87

Sumário

1	Introdução	14
1.1	<i>Objetivos</i>	16
1.2	<i>Materiais e Métodos</i>	17
1.3	<i>Organização do documento</i>	18
2	Conceitos Fundamentais	20
2.1	<i>Biologia molecular e evolutiva</i>	20
2.2	<i>Análise filogenética computacional</i>	24
3	Trabalhos Correlatos	31
3.1	<i>Revisão Sistemática de ferramentas de análise filogenética</i>	31
3.1.1	Conceitos teóricos	32
3.1.2	Categorização	32
3.1.3	Considerações finais da revisão sistemática	39
3.2	<i>Análises considerando genes exclusivos</i>	40
3.3	<i>Análises considerando bactérias do gênero Xanthomonas</i>	41
4	Abordagem para Análise Filogenética de Genes Exclusivos . .	44
4.1	<i>Representação dos dados</i>	47
4.1.1	Arquivo de alinhamento - M8	47
4.1.2	Matriz de Presença	49
4.1.3	FASTA	50
4.2	<i>Identificação dos Potenciais Genes Exclusivos</i>	50
4.3	<i>Identificação de erros de anotação e genes realmente exclusivos</i>	53
4.4	<i>Identificação de famílias de genes no próprio genoma</i>	57
4.5	<i>Taxonomia de genes exclusivos</i>	58
5	Experimentos e Resultados	62
5.1	<i>Descrição dos Dados</i>	62
5.2	<i>Estudo de Caso</i>	66
5.3	<i>Descrição dos Experimentos</i>	68
5.4	<i>Análise de Erros de Anotação</i>	69

5.5	<i>Análise de alinhamentos com outras famílias de genes no próprio genoma</i>	80
5.6	<i>Análise de alinhamentos dos genes exclusivos em outras espécies . . .</i>	85
6	Conclusão	91
6.1	<i>Principais contribuições</i>	93
6.2	<i>Trabalhos Futuros</i>	94
	Referências¹	95
	Apêndice A – Protocolo para Revisão Sistemática: Estado da Arte em análise filogenética de um gene específico	100

¹ De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

1 Introdução

Atualmente, existem incontáveis organismos vivos presentes em diversos grupos (entidades, táxons); cada qual apresentando características diferentes. A combinação dessas características com os grupos em que os organismos estão inseridos forma o que conhecemos como diversidade biológica. A diversidade biológica, ou biodiversidade, é a variabilidade de organismos vivos de todas as origens, compreendendo ecossistemas terrestres e marinhos, além da diversidade dentro de uma espécie e entre espécies (MATIOLI, 2001). Organizando o conhecimento existente atualmente sobre diversidade biológica, é possível encontrar trilhões de características genéticas cruzadas com trilhões de grupos. Isso representa, minimamente, a base de dados da diversidade biológica existente atualmente (SETUBAL; MEIDANIS, 1997).

A diversidade biológica atual reflete a evolução das espécies. Pode-se considerar que duas espécies atuais quaisquer possuem, pelo menos, uma espécie ancestral comum. Para três espécies atuais quaisquer, duas possuem um mesmo ancestral, que não é comum à terceira. Aplicando essa lógica em todas as espécies existentes hoje, é possível encontrar uma sequência de divisões desde o primeiro ancestral (AMORIM, 2002; MATIOLI, 2001). O estudo dessa sequência de divisões entre grupos de organismos, ou seja, da relação evolutiva, é chamado de filogenia. Isto posto, temos que a diversidade biológica atual pode ser considerada um recorte temporal da filogenia, ou seja, ela não representa completamente a ancestralidade das espécies (MIYAMOTO; CRACRAFT, 1991).

A análise filogenética busca examinar a relação evolutiva de cada espécie considerando suas características genéticas. Essa relação pode ser identificada através de comparações de genes, ou grupos de genes, de determinada espécie com o genoma das demais espécies existentes (MIYAMOTO; CRACRAFT, 1991). Isso é feito, tipicamente, comparando um determinado gene com os genes do *National Center for Biotechnology Information gene and genome database* (NCBI), a fim de encontrar todos os organismos conhecidos envolvidos na filogenia desse gene para esta espécie. A partir dessas comparações é possível gerar uma árvore filogenética. Esse tipo de visualização da filogenia auxilia o trabalho de pesquisadores da área de biologia comparativa. A identificação da origem de um gene é de grande importância para a área de bioinformática e biologia comparativa, pois diversos problemas atuais requerem estudos filogenéticos.

Porém, tal processo é muito complicado de ser feito sem o auxílio de ferramentas computacionais robustas, devido ao grande volume de dados envolvidos e pela dificuldade em identificar ancestrais comuns, dado que tais espécies não foram sequenciadas. Por conta disso, soluções robustas que integrem e tragam melhorias às soluções já existentes irão facilitar o trabalho de pesquisadores da área.

Existem diversas ferramentas computacionais que realizam etapas do processo de análise filogenética; como ferramentas que realizam a montagem de sequências, identificação e anotação de genes, alinhamento de genomas, comparação de genes e construção de árvores filogenéticas. As ferramentas existentes atualmente são capazes de apresentar resultados detalhados, quando o volume de dados é pequeno. Porém, muitas vezes os dados de genomas sequenciados públicos podem conter inconsistências ou potenciais erros de anotação (marcações incorretas de sequências como genes), o que causa problemas na interpretação das filogenias geradas a partir desses genes mal anotados. Além disso, tais ferramentas encontram dificuldades ao lidar com volumes maiores de dados, onde existe cruzamento de informações de espécies muito distantes (ou muito próximas) evolutivamente, e muitas especificidades de espécies. Por fim, outro problema com o qual as ferramentas existentes têm dificuldade de lidar é descobrir origens de genes exclusivos das espécies.

Genes exclusivos são genes encontrados em apenas uma espécie, quando é feita uma análise considerados apenas grupos de espécies próximas, mas que não são necessariamente característicos da espécie (ou gênero ou família) do organismo em questão. A identificação de genes exclusivos de alguma espécie (ou cepa) pode ser fundamental para entender as características específicas dessa espécie e pode servir de marcador molecular para a correta identificação de, por exemplo, uma doença, de forma a permitir o uso do tratamento mais específico e adequado ou pode ser útil para o desenvolvimento de vacinas (MOREIRA et al., 2010).

Foi a importância da descoberta de filogenias de genes exclusivos e a dificuldade de garantir a consistência nas anotações genéticas que motivaram este trabalho que motivaram este trabalho. O desenvolvimento ocorreu com base em hipóteses identificadas ao longo dos estudos iniciais. Além disso, foram realizados experimentos para analisar as hipóteses identificadas e para validar a estratégia implementada pelos algoritmos. Nos experimentos realizados, foram utilizados genes exclusivos de bactérias do gênero *Xanthomonas*, que possui bactérias que causam grandes prejuízos econômicos pois causam doenças em diversas espécies de plantas.

Isto posto, o presente projeto visa trabalhar com genes exclusivos, com o objetivo de especificar e implementar uma abordagem suficientemente robusta que integre as tarefas do processo de análise filogenética. Este projeto utiliza ferramentas consolidadas na área de análise filogenética e biologia comparativa para tarefas gerais do processo (como alinhamento e comparação de genes e sequências de nucleotídeos). Além disso, são aplicadas melhorias em ferramentas desenvolvidas pelo grupo de pesquisa, a fim de torná-las capazes de lidar com a identificação e análise de genes exclusivos. Por fim, a tarefa principal deste projeto consistiu-se no desenvolvimento de ferramentas integradas para lidar com problemas de inconsistências em dados genéticos, e filogenia de genes exclusivos. Todas essas atividades visam a identificação de possíveis origens desses genes exclusivos.

Este trabalho está focado apenas na análise de procariontes e o estudo de caso foi realizado considerando apenas bactérias. Adicionalmente, a palavra gene será utilizada muitas vezes de maneira intercalada com sequência codificante ou CDS (*coding DNA sequence*).

1.1 Objetivos

O objetivo geral deste trabalho é desenvolver uma estratégia, implementando ferramentas computacionais, para lidar com o problema de análise filogenética de espécies a partir de genes exclusivos de genomas próximos e identificar possíveis origens de tais genes. Para atingir este objetivo, foram realizadas comparações entre genomas de bactérias (levando em consideração seus genes e sequências de nucleotídeos de seus cromossomos) do banco de dados de genomas do NCBI, utilizando ferramentas consolidadas da área de bioinformática, para dar suporte ao desenvolvimento da abordagem.

Nesse contexto, os seguintes objetivos específicos foram estabelecidos:

- Levantamento do estado da arte de ferramentas de análise filogenética;
- Estudo do estado da arte de análises filogenéticas de bactérias do gênero *Xanthomonas* e genes exclusivos;
- Avaliação da adequabilidade de abordagens de análise filogenética no problema de genes exclusivos;
- Desenvolvimento de abordagem robusta, que envolva todas as etapas do processo de análise filogenética, para tratar o problema específico de genes exclusivos;

- Análise dos resultados obtidos pelas estratégias propostas.

1.2 Materiais e Métodos

Esta seção apresenta, em linhas gerais, a metodologia utilizada para desenvolvimento deste trabalho a fim de atingir os objetivos propostos.

Inicialmente, para realizar a revisão dos conceitos fundamentais de biologia molecular e evolutiva e sobre teorias e técnicas computacionais sobre análise filogenética, foram feitas revisões bibliográficas baseadas na leitura de livros e artigos clássicos dessas áreas. Também foram feitos levantamentos bibliográficos de trabalhos sobre bactérias do gênero *Xanthomonas* e estudos com genes exclusivos, a fim de entender a importância de se estudar tais assuntos e saber o que tem sido feito nessas áreas.

Já para realizar o levantamento do estado da arte, foi utilizada a metodologia de revisão sistemática proposta por (KITCHENHAM, 2004). Segundo Kitchenham (2004), a revisão sistemática consiste em uma metodologia para levantamento bibliográfico que identifica, avalia e interpreta pesquisas relevantes dentro de uma determinada área ou objeto de pesquisa. O objetivo dessa metodologia é organizar a revisão, de forma que seja possível documentar o processo de levantamento bibliográfico, a fim de criar revisões parametrizadas e reproduzíveis.

A revisão sistemática realizada durante o desenvolvimento deste trabalho foi feita com o objetivo de entender o estado da arte de técnicas que envolvem o processo de análise filogenética, identificando as ferramentas mais utilizadas neste processo.

Em seguida, técnicas e métodos foram escolhidos para serem utilizados no problema de análise filogenética de genes exclusivos. Também foram escolhidas ferramentas já consolidadas na área para serem utilizadas no trabalho para alinhamento de genomas.

A fase de levantamento de dados envolveu a escolha de espécies a serem estudadas. Foram escolhidos 15 genomas de bactérias do gênero *Xanthomonas*. Esses genomas foram obtidos do banco de dados biológicos do NCBI (*National Center for Biotechnology Information gene and genome database*), o GenBank¹, representados por sequências de nucleotídeos em arquivos no formato FASTA.

¹ <http://www.ncbi.nlm.nih.gov/genbank>

Os arquivos FASTA serviram como entrada para a ferramenta BLASTP², que considera sequências que codificam proteínas e BLASTN³, para qualquer sequência de nucleotídeos. Essas ferramentas realizam alinhamentos de sequências. Os alinhamentos são a base dos experimentos desenvolvidos neste trabalho, além de servirem como entrada para outras ferramentas auxiliares utilizadas ao longo deste trabalho, desenvolvidas pelo grupo de pesquisa.

Foi realizado um estudo de caso com o genoma de uma das espécies para validar a viabilidade e relevância da abordagem. Nesse estudo de caso, surgiram hipóteses de análises e origem dos genes exclusivos que guiaram os desenvolvimentos e experimentos realizados.

A seguir, foram implementadas as ferramentas que tinham como objetivo validar as hipóteses identificadas. Tais ferramentas foram desenvolvidas em Java. Foram realizados três experimentos, onde cada um tinha o objetivo de validar uma das hipóteses. A análise dos resultados dos experimentos envolveu o desenvolvimento de scripts simples em Python e em R para tratamento de arquivos de resultados e visualização de dados.

Esta seção apresentou as etapas desenvolvidas neste trabalho de forma resumida. Uma descrição mais detalhada sobre os dados e ferramentas implementadas (incluindo o estudo de caso) pode ser obtida no capítulo 4. A descrição dos experimentos e resultados encontrados (incluindo os resultados do estudo de caso) está descrita com maiores detalhes no capítulo 5.

1.3 Organização do documento

Esta dissertação de mestrado está dividida em 6 capítulos, incluindo este capítulo introdutório. Os capítulos seguintes estão organizados da seguinte forma:

- O capítulo 2 apresenta um resumo sobre os conceitos fundamentais e teóricos de biologia molecular e evolutiva, bem como conceitos de biologia computacional e bioinformática, com base nos quais foi feita a elaboração deste trabalho.
- O capítulo 3 apresenta a revisão de literatura correlata ao trabalho apresentado nesta dissertação, pesquisados durante a revisão sistemática e revisão bibliográfica desenvolvidas. Esses trabalhos dizem respeito, principalmente, a ferramentas de

² <https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins>

³ <https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Nucleotides>

bioinformática mais utilizadas no processo de análise filogenética, bem como envolvem estudos de genes exclusivos e de bactérias do gênero *Xanthomonas*.

- O capítulo 4 apresenta a descrição das etapas desenvolvidas neste trabalho para a abordagem de análise filogenética de genes exclusivos. Essa descrição contém aspectos sobre o fluxo implementado para cada etapa de análise filogenética realizada.
- O capítulo 5 apresenta a descrição do conjunto de dados utilizados nos experimentos de validação da abordagem de análise filogenética de genes exclusivos, também descreve os experimentos realizados, apresenta os principais resultados obtidos a partir desses experimentos e algumas análises sobre esses resultados.
- Por fim, no capítulo 6 são apresentadas considerações finais deste trabalho, incluindo as principais contribuições do estudo e possibilidades de extensão do mesmo.

2 Conceitos Fundamentais

O presente capítulo apresenta conceitos fundamentais necessários para o pleno entendimento do trabalho descrito nas seções seguintes desta dissertação. Tais conceitos envolvem desde conceitos relacionados à biologia molecular e evolutiva, até conceitos e ferramentas computacionais usadas neste projeto, passando por conceitos de biologia comparativa e bioinformática.

2.1 *Biologia molecular e evolutiva*

As formas de vida mais antigas e simples datam de 3,5 bilhões de anos atrás (descobertas a partir de células fossilizadas), tendo se originado em condições hostis (MATIOLI, 2001). Tais formas de vida constituem os ancestrais diretos dos seres vivos atuais. Por meio de um processo lento de transformação, chamado evolução, tais formas de vida evoluíram e se diversificaram, gerando os organismos complexos existentes atualmente, assim como outros organismos simples (SETUBAL; MEIDANIS, 1997).

Todos os organismos vivos, sejam eles simples ou complexos, compartilham os mesmos compostos químicos: proteínas e ácidos nucleicos. Proteínas são responsáveis pelo que os seres vivos são fisicamente; enquanto ácidos nucleicos (especialmente o ácido desoxirribonucleico (DNA)) são responsáveis por codificar as informações necessárias para criação das proteínas e por repassar tais informações para as próximas gerações de organismos (SETUBAL; MEIDANIS, 1997).

Visto que o foco do presente trabalho está na identificação de filogenias (explicada mais adiante neste capítulo) e na análise da evolução de características de seres vivos, moléculas de DNA serão descritas mais detalhadamente.

A molécula de DNA é formada por duas cadeias de moléculas (chamadas de fitas). Cada fita de DNA é constituída por unidades básicas chamadas nucleotídeos, que por sua vez, é composta de açúcar, fosfato e uma base nitrogenada. Existem quatro bases possíveis na molécula de DNA: adenina(A), guanina(G), citosina(C) e timina(T). Cada fita da molécula de DNA se conecta à outra, através das suas bases: a base A se conecta à base T e a base C se conecta à base G. Dessa forma A e T são ditas bases complementares, assim como as bases C e G (SETUBAL; MEIDANIS, 1997).

Como já dito, a molécula de DNA possui trechos que contêm informações para formação de proteínas ou moléculas de RNA. Tais trechos são conhecidos como genes. Ou seja, um gene é um trecho de DNA que contém sequências que não são traduzidas em proteínas (íntrons) e sequências codificadoras (éxons) que contêm a informação necessária para realizar a síntese de proteínas ou moléculas de RNA. O tamanho dos genes é variável, podendo chegar a ter mais de 10.000 bases no genoma humano, por exemplo (SETUBAL; MEIDANIS, 1997).

Moléculas de proteínas são formadas por sub-moléculas de aminoácidos. Dessa forma, para especificar uma proteína, é necessário especificar cada aminoácido que ela contém. Essa é a tarefa do DNA em um gene, utilizando trincas de nucleotídeos para especificar cada aminoácido. Cada trinca de nucleotídeo é chamada de códon; alguns códons codificam o mesmo aminoácido (SETUBAL; MEIDANIS, 1997; ZAHA; FERREIRA; PASSAGLIA, 2003).

Existem 3 sequências de bases (códon) que não correspondem a nenhum aminoácido: TAA, TGA, TAG. Esses são chamados códons de parada (*stop codons*) e são utilizados para identificar o fim da leitura de uma sequência de bases que codifica uma proteína (MATIOLI, 2001).

Cada molécula de DNA contém milhares de genes e é chamada de cromossomo. O conjunto de cromossomos de um organismo é chamado de genoma, que constitui a informação genética desse organismo, necessária para sua sobrevivência e reprodução (ZAHA; FERREIRA; PASSAGLIA, 2003). O número de cromossomos em um genoma é uma característica de uma espécie, sendo que o genoma humano, por exemplo, é composto por 46 cromossomos, enquanto em um rato o número é 40. Dentre os diversos organismos existentes atualmente, os cromossomos das bactérias são os mais utilizados em pesquisas genéticas. Isso ocorre, porque bactérias são organismos unicelulares que possuem apenas um cromossomo, portanto, seu genoma costuma ser menor do que dos outros organismos. Além disso, bactérias podem se reproduzir por simples replicação de DNA em um curto período de tempo (SETUBAL; MEIDANIS, 1997).

Os genomas dos organismos vivos definem as características fenotípicas de cada espécie. Algumas espécies compartilham as mesmas características que tiveram origem a partir de um ancestral comum. Esse fenômeno é chamado de homologia (MATIOLI, 2001).

O conceito de homologia pode ser transferido para a biologia molecular. Dois ou mais genes são homólogos quando suas sequências de DNA derivam de um mesmo gene

ancestral comum, podendo ou não ter a mesma função. Genes homólogos podem surgir de formas distintas: por evento de especiação ou por duplicação genética (FITCH, 1970).

Quando os genes homólogos surgem por especiação são chamados de ortólogos. Nesse caso, cada cópia do gene divergiu para espécies distintas, descendendo de um gene presente no último ancestral comum das espécies, apresentando sequência de nucleotídeos semelhantes e, tipicamente, codificando proteínas similares com função correspondente. Genes ortólogos são importantes na determinação da função de novos genes, pois a função tende a ser a mesma (FITCH, 1970; FREEMAN; HERRON, 2009).

Já os genes homólogos que surgem por duplicação genética dentro do genoma de uma mesma espécie são chamados de genes parálogos. Nesse caso, um gene é duplicado dentro de um genoma, de modo que as cópias divergem do gene original através de mutações e podem ter funções distintas da original. Genes parálogos são úteis no estudo da evolução das espécies (MATIOLI, 2001; ZAHA; FERREIRA; PASSAGLIA, 2003).

A figura 1 ilustra os conceitos de genes ortólogos e parálogos. Uma espécie ancestral, que possui um gene A, sofre um evento de especiação e duas novas espécies surgem, com os genes B e C descendentes do gene ancestral A respectivamente. Os genes B e C são ortólogos. A seguir, ocorre uma duplicação genética na espécie que contém o gene B e gera uma cópia desse gene, passando a ter os genes B e D dentro do genoma da própria espécie. Os genes B e D são parálogos. Além disso, todos eles continuam sendo homólogos entre si.

Figura 1 – Conceitos de genes homólogos, ortólogos e parálogos.

Fonte: Priscilla Koch Wagner, 2018

Genes homólogos podem ainda surgir a partir de transferência horizontal (ou lateral). Nesse caso, são chamados de genes xenólogos (HILLIS; MORITZ; MABLE, 1996). A transferência horizontal, diferente da transferência vertical, onde os genes são passados através de divisão celular, é um processo no qual partes de DNA são transferidas entre organismos de diferentes espécies (ZAHA; FERREIRA; PASSAGLIA, 2003).

Esse fenômeno é tipicamente observado em procariotos. Um exemplo de transferência horizontal acontece nas bactérias que possuem plasmídeos. Plasmídeos são moléculas de DNA independentes que são capazes de se auto-replicar. Seu DNA é independente do cromossomo das bactérias. Dessa forma, bactérias que possuem plasmídeos, podem absorver parte do DNA do plasmídeo e incorporá-lo ao seu próprio DNA (ZAHA; FERREIRA; PASSAGLIA, 2003).

A transferência horizontal pode trazer importantes implicações evolutivas. Possibilita, por exemplo, que determinada bactéria adquira um gene que a torne resistente à antibióticos (OCHMAN; LAWRENCE; GROISMAN, 2000).

Apesar da importância evolutiva, inferências filogenéticas baseadas em um único gene podem ser prejudicadas, pois se houver transferência horizontal do gene de referência entre organismos distantes evolutivamente, a análise filogenética pode acabar por aproximá-los.

Dentre as inferências filogenéticas que podem ser dificultadas no processo de transferência horizontal estão aquelas baseadas em genes exclusivos. Um gene é considerado exclusivo quando é encontrado em um ou alguns organismos, em uma análise que considera um grupo de espécies (ou sub-espécies próximas), mas que não são característicos da espécie (ou do gênero ou da família) do organismo em questão. Vale ressaltar, portanto, que ao modificar o escopo da análise, adicionando ou removendo espécies (ou gêneros ou famílias) consideradas em um estudo de biologia comparativa, os genes considerados exclusivos possivelmente serão alterados, dado que podem existir cópias similares a genes exclusivos nos novos genomas adicionados; ou podem ser identificados novos genes exclusivos; ou ainda, genes identificados como homólogos podem passar a ser considerados exclusivos, se o genoma que possuía o gene semelhante for removido da análise. A identificação de genes exclusivos e de sua origem é de grande importância na área de biologia comparativa (MOREIRA et al., 2010), podendo ser fundamental para entender as características específicas de uma espécie, contrapondo às outras espécies próximas da análise, além de servir de marcador molecular para a correta identificação de, por exemplo, uma doença, de

forma a permitir o uso do tratamento mais específico e adequado; ou ainda ser utilizado no desenvolvimento de vacinas (LIMA-JUNIOR et al., 2016).

Neste trabalho, foram utilizados genes exclusivos de bactérias do gênero *Xanthomonas* para estudo e validação da abordagem proposta. O gênero *Xanthomonas* engloba um grande grupo de bactérias fitopatogênicas, ou seja, que causam doenças em diversas espécies vegetais. Essas bactérias afetam diversas partes das plantas, tais como folhas, caule, sementes, frutos. Alguns sintomas comuns são lesões e manchas nas folhas, murchidão e podridão, hipertrofia (excesso de crescimento) e até surgimento de cancrios (RUDOLPH, 1993).

Por atacar e causar danos severos em diversas espécies vegetais, as infecções causadas por bactérias desse gênero resultam em perdas de colheitas inteiras e, conseqüentemente, grandes prejuízos econômicos para a agricultura (RYAN et al., 2011; JALAN et al., 2013). Por conta disso, as *Xanthomonas* vem sendo amplamente estudadas pela comunidade científica, visto que estudos relacionados a essas bactérias podem prover um melhor entendimento sobre as patogenias e sua evolução, a fim de colaborar no tratamento e prevenir o surgimento de novas doenças (ZHANG et al., 2015).

Atualmente existem aproximadamente 400 sequências genômicas de *Xanthomonas* sequenciadas no GenBank, sendo que 50 correspondem a genomas completamente sequenciadas. Porém, grande parte dessas sequências consiste de fragmentos com qualidade variada de sequenciamento e com possíveis erros no processo de anotação (JACQUES et al., 2016).

2.2 Análise filogenética computacional

Para a realização de estudos científicos de um genoma, é necessário, primeiramente, extrair a informação básica da molécula de DNA: sua sequência de pares de bases. O processo que identifica e obtém a ordem correta dessas informações é chamado de sequenciamento (MIYAMOTO; CRACRAFT, 1991; SETUBAL; MEIDANIS, 1997).

Apesar dos avanços tecnológicos que ocorreram nos últimos anos no processo de sequenciamento terem proporcionado melhorias em relação ao tamanho e quantidade de sequências processadas e em relação ao tempo de processamento, ainda não é possível realizar o sequenciamento do genoma por completo de uma única vez. Os sequenciadores

mais poderosos são capazes de sequenciar trechos com comprimento de aproximadamente 700 pares de bases; enquanto um cromossomo de DNA humano, por exemplo, possui cerca de 10^8 pares de bases. Desse modo, fragmentos de moléculas de DNA são isolados e sequenciados separadamente. A seguir, é necessário realizar a montagem dos fragmentos da sequência para identificar a sequência original (SETUBAL; MEIDANIS, 1997; FREEMAN; HERRON, 2009).

Um dos métodos mais utilizados para realizar o sequenciamento de cadeias longas é o método *shotgun*. Esse método consiste em dividir a molécula de DNA aleatoriamente em vários segmentos pequenos, possíveis de ser sequenciados. Como a divisão é aleatória, tem-se diversas subsequências sobrepostas. Essas subsequências são então re-ordenadas e remontadas, através de programas computacionais baseados na sobreposição dos fragmentos obtidos. No final do processo de montagem espera-se obter uma sequência única para genomas que possuem um único cromossomo, ou uma sequência por cromossomo para representar o genoma de indivíduos com mais de um cromossomo (SETUBAL; MEIDANIS, 1997).

O processo de sequenciamento e montagem costumam lidar com diversos problemas e erros que vão desde material genético contaminado, que pode possuir trechos de sequência não pertencentes ao organismo de interesse, até genomas com muitas regiões repetidas ou complexas, visto que há dificuldade em realizar a correta ordenação dessas regiões. A sequência final ainda pode ser comprometida caso as sub-sequências apresentem baixa cobertura em relação à sequência original. Dados esses problemas enfrentados no processo, identifica-se uma grande dificuldade em se obter uma sequência única que represente um genoma completo, visto que a maioria dos genomas sequenciados e montados atualmente ainda estão incompletos ou não finalizados (MOREIRA, 2015).

O processo de sequenciamento e montagem de sequências é de extrema importância, pois permite obter informações relevantes sobre os genomas de interesse, tais como estrutura e função de genes, identificação de genes obtidos por transferência lateral e relações evolutivas. Este trabalho não explora o processo de sequenciamento e montagem e seus potenciais problemas.

Após a realização do sequenciamento e montagem do genoma, pode ser realizada a anotação de genes.

O processo de anotação visa a identificar quais sequências de nucleotídeos do genoma sequenciado tem a informação genética necessária para codificar uma proteína, ou

seja, quais sequências de nucleotídeos são genes. Esse processo busca também atribuir aos genes identificados nas sequências o máximo de informação biológica possível, tal como a atribuição de suas prováveis funções a fim de entender fenômenos biológicos que ocorrem no organismo ou na espécie. A anotação de uma sequência é um processo complexo e pode atribuir diversos atributos à uma sequência. Neste trabalho, o conceito de anotação utilizado se resume à marcação de uma sequência como gene ou não.

O processo de anotação pode ser realizado automaticamente, por meio de programas computacionais capazes de anotar um conjunto de genes de um organismo de uma só vez; ou manualmente, realizada por um anotador para cada gene separadamente. É possível ainda, realizar o processo de anotação de genes em laboratório, identificando quais sequências de nucleotídeos se expressam como gene, ou seja, quais sequências codificam proteínas. Os processos manuais tendem a ser realizados com mais cuidado e, portanto, são considerados mais confiáveis que a anotação automática. Porém, por ser um processo mais caro e demorado e por conta do grande volume de dados, a anotação automática é a mais utilizada (SETUBAL; MEIDANIS, 1997).

Como esse processo é muito passível de erros, parte significativa das sequências depositadas em bancos de dados públicos de genomas pode estar erroneamente anotada e, ainda assim, serve de base para estudos de genômica, análise filogenética e outras áreas. Desse modo, a propagação de erros de anotação é um dos maiores problemas que esse processo gera, a qual afeta outros processos que se utilizam direta e indiretamente dos resultados gerados por ele (MOREIRA, 2015).

A partir das sequências de nucleotídeos montadas ou genes anotados é possível realizar comparação de sequências ou de genes (sequências anotadas).

Comparação de sequências é uma das operações mais importantes em biologia computacional, pois serve de base para diversas outras manipulações mais complexas. Visa a representar uma hipótese evolutiva das sequências, visto que pode identificar um determinado grau de similaridade entre sequências distintas que compartilham um ancestral comum, estando relacionadas evolutivamente. Em resumo, o processo de comparação de genes ou genomas consiste em alinhar sequências de DNA, a fim de encontrar similaridades e divergências entre as diferentes sequências. Embora pareça uma tarefa simples, a comparação de sequências envolve técnicas computacionais complexas (MIYAMOTO; CRACRAFT, 1991; SETUBAL; MEIDANIS, 1997).

A tarefa de comparação de sequências envolve dois conceitos principais: similaridade de duas sequências, medida de quão parecidas são as sequências; e alinhamento, onde as sequências são colocadas “umas sobre as outras” a fim de obter a melhor correspondência entre as posições/caracteres similares de duas ou mais sequências de nucleotídeos (SETUBAL; MEIDANIS, 1997; FREEMAN; HERRON, 2009).

As sequências em um alinhamento podem ter tamanhos diferentes. Para realizar o alinhamento entre as sequências e para que seja possível visualizar similaridade com maior facilidade, garante-se que elas tenham o mesmo comprimento inserindo espaços (lacunas) em locais arbitrários ao longo das sequências a serem alinhadas. As lacunas correspondem a eventos de inserção ou deleção de nucleotídeos e são representadas por um hífen(-) ou ponto(.) nas sequências. Após garantir o mesmo tamanho, as sequências são posicionadas uma sobre a outra, criando uma correspondência entre caracteres ou lacunas na primeira sequência e caracteres ou lacunas na segunda. Adicionalmente, vale ressaltar que nenhuma lacuna da primeira sequência deve alinhar com lacunas da segunda (SETUBAL; MEIDANIS, 1997). A figura 2 apresenta um exemplo de alinhamento de sequências.

Figura 2 – Alinhamento entre quatro sequências de nucleotídeos. A parte *a* apresenta as sequências originais antes do alinhamento. A parte *b* apresenta o alinhamento das sequências, incluindo as lacunas adicionadas (hífens).

Fonte: Verli (2014).

Algoritmos de alinhamento buscam identificar o alinhamento que minimize as diferenças entre sequências. A similaridade entre elas é calculada a partir da identidade do alinhamento, ou seja, porcentagem de caracteres idênticos e nas mesmas posições da sequência.

Algoritmos de alinhamento podem ser classificados de acordo com a estratégia aplicada para comparação, sendo categorizados em globais ou locais. No alinhamento global, todos os nucleotídeos das sequências são alinhados uns com os outros na extensão completa das sequências. Já no alinhamento local, as regiões que apresentarem maior identidade serão alinhadas (SETUBAL; MEIDANIS, 1997; VERLI, 2014).

Os alinhamentos podem ainda ser realizados apenas entre duas sequências, nesse caso é chamado de alinhamento par-a-par (*pairwise alignment*). Pode ainda ser realizado entre três ou mais sequências, sendo chamado de alinhamento múltiplo (*multiple sequence alignment*). Vale ressaltar que, quanto mais sequências fizerem parte do alinhamento, mais tempo será demandado para processá-lo e maior será a dificuldade para obter um alinhamento ótimo.

A ferramenta para alinhamento e comparação genética mais utilizada é o BLAST. BLAST é a sigla para o termo em inglês *Basic Local Alignment Search Tool*. É uma ferramenta de alinhamento, muito utilizada por pesquisadores de filogenética e biologia comparativa e evolutiva, que busca regiões similares em sequências de genomas. Para isso, utiliza comparação de sequência de nucleotídeos ou aminoácidos com sequências presentes nas bases de genomas e identifica correspondências relevantes entre as sequências por meios de inferência estatística (ALTSCHUL et al., 1990).

Pode ser utilizada para identificação de relações evolutivas, inferência funcional, além de auxiliar na identificação de membros de famílias de genes. O resultado dos alinhamentos locais do BLAST é obtido a partir da computação de uma matriz de pontuações para todos os possíveis pares de segmentos formados pela sequência fornecida com todas as sequências presentes no banco de dados. O objetivo é encontrar os pares de segmentos que formam um par de segmento máximo, ou seja, o par com pontuação máxima para um dado comprimento. A pontuação define uma medida de similaridade local para o par de sequências (ALTSCHUL et al., 1990).

O resultado do processamento dos alinhamentos do BLAST é um arquivo contendo 12 campos com informações sobre cada alinhamento encontrado, sendo eles: *query*, *subject*, *percent identities*, *alignment length*, *mismatches*, *gap openings*, *query sequence start*, *query sequence end*, *subject sequence start*, *subject sequence end*, *e-value*, *bit-score*. A descrição desses doze campos está detalhada na seção 4.1 do capítulo 4 deste trabalho.

Ao longo dos anos foram criadas algumas variações da ferramenta. Por exemplo o TBLASTN, que realiza o alinhamento de aminoácidos com sequências de nucleotídeos

da base de dados. Isso é feito através da tradução das sequências de nucleotídeos em hipotéticas sequências de aminoácidos, e essas sequências são alinhadas com a sequência de entrada (GERTZ et al., 2006).

Como já dito, comparação ou alinhamento de sequências é uma tarefa que serve como base para outras aplicações da biologia computacional, tais como análise de padrões em sequências, identificação de sequências homólogas, montagem de genomas, estudos de inferência estrutural e funcional de proteínas (VERLI, 2014; MOREIRA, 2015)

A tarefa de comparação de sequências homólogas é a tarefa básica na análise da ancestralidade das espécies, ou seja, espécies podem ser classificadas através da análise filogenética, com base em genes homólogos. A filogenética é uma sub-área da biologia evolutiva que visa a estudar e explicar a história evolutiva de espécies, populações, gêneros ou outras unidades taxonômicas, além de identificar relações entre espécies com base em ancestrais comuns (SETUBAL; MEIDANIS, 1997).

A filogenética busca reconstruir relações evolutivas entre organismos existentes hoje e organismos já extintos, considerando que compartilham uma origem evolutiva em comum. Identificar fatores importantes no processo evolutivo das espécies é um dos problemas centrais da biologia comparativa atualmente (SETUBAL; MEIDANIS, 1997; MOREIRA, 2015).

Para representar visualmente a história evolutiva dos organismos, comumente são utilizadas árvores filogenéticas, ou filogenias. Nelas, é possível observar relações entre os organismos e seus ancestrais, além de avaliar a proximidade evolutiva de uns em relação aos outros. A árvore filogenética é um diagrama ramificado no qual os nós terminais representam espécies atuais que são conectadas através de nós internos, que representam seus ancestrais comuns (SETUBAL; MEIDANIS, 1997; MOREIRA, 2015).

Os nós terminais podem também ser chamados de unidades taxonômicas operacionais (*operational taxonomical units* - OTUs). Podem representar táxons (grupos de organismos, por exemplo, espécie, gêneros, famílias) ou indivíduos únicos, podendo ser de espécies diferentes ou subespécies que se subdividiram a partir de uma espécie, sendo diferenciadas com base em variações fenotípicas ou por formarem diferentes grupos distinguidos por variações genéticas, ou ainda, no caso de bactérias, podem ser subconjuntos de espécies (SCHAECHTER, 2009; BRENNER; STALEY; KRIEG, 2005).

Uma árvore filogenética registra a sequência em que as linhagens apareceram ao longo do tempo. Além disso, identifica os organismos que são evolutivamente mais próximos

ou mais distantes, de acordo com um ancestral comum compartilhado (MIYAMOTO; CRACRAFT, 1991; SETUBAL; MEIDANIS, 1997; FREEMAN; HERRON, 2009).

Um exemplo de árvore filogenética é apresentado na figura 3. Segundo este exemplo, seres humanos e chimpanzés são geneticamente mais próximos uns dos outros do que em relação aos outros primatas, ou seja, apresentam um ancestral em comum mais próximo. Uma relação parecida é apresentada entre o gibão e o gibão preto.

Figura 3 – Árvore filogenética de primatas: gibão preto (siamang), gibão (gibbon), orangotango (orangutan), gorila (gorilla), humano (human), chimpanzé (chimpanzee).

Fonte: Setubal e Meidanis (1997), p.176.

Pesquisadores de biologia comparativa utilizam árvores filogenéticas com o objetivo de estudar duas características principais da história evolutiva dos organismos, que são representadas por elas. Uma delas é a topologia, que indica como os nós interiores se conectam uns aos outros e aos nós terminais. A outra característica é a distância entre pares de nós, que estima a distância evolutiva entre os indivíduos representados pelos nós (SETUBAL; MEIDANIS, 1997).

Um problema existente na geração de árvores filogenéticas é que geralmente não existem dados suficientes sobre os ancestrais mais distantes das espécies atuais. Mesmo quando existem dados suficientes, não é possível afirmar com certeza que determinado fóssil é da mesma espécie que o ancestral das espécies atuais. Por conta disso, a história evolutiva muitas vezes é inferida com base nos dados comparativos entre as espécies atuais (SETUBAL; MEIDANIS, 1997). Dessa forma, a árvore apresentada na figura 3 é uma hipótese da árvore evolutiva de primatas.

3 Trabalhos Correlatos

Neste trabalho, ferramentas que implementam uma abordagem para análise de anotação genética, origem de genes exclusivos e identificação de famílias de genes próximas forem implementadas considerando genes exclusivos de espécies do gênero *Xanthomonas*. Por esse motivo, viu-se a necessidade de identificar trabalhos nessas frentes que vem sendo desenvolvidos pela comunidade científica.

Durante o desenvolvimento deste trabalho foi realizada uma revisão sistemática sobre ferramentas de análise filogenética que estão sendo utilizadas ou desenvolvidas nos últimos anos. Além disso, identificou-se a importância de identificar trabalhos que estudam análise filogenética focada em genes exclusivos; além de estudos utilizando bactérias do gênero *Xanthomonas*.

Este capítulo apresenta os trabalhos relacionados com o projeto desta dissertação. A seção 3.1 apresenta os resultados da revisão sistemática desenvolvida; a seção 3.2 apresenta alguns trabalhos que analisam genes exclusivos das espécies; e, a seção 3.3 apresenta estudos utilizando bactérias do gênero *Xanthomonas*.

3.1 Revisão Sistemática de ferramentas de análise filogenética

Com o objetivo de contextualizar este trabalho em relação aos métodos e técnicas que vêm sendo utilizados na literatura para a realização do processo de análise filogenética ou parte dele, foi realizada uma revisão sistemática dos trabalhos da área. Ao final da revisão sistemática é possível visualizar, de forma abrangente, o estado da arte na área de análise filogenética baseada em um determinado gene. A revisão sistemática desenvolvida neste projeto adotou uma mescla das metodologias propostas por Kitchenham (2004). Um resumo da revisão sistemática realizada é apresentada nesta seção.

Durante a condução da revisão foram identificados 48 artigos, definida pelo protocolo apresentado no Apêndice A, nas bases de dados de artigos: ACM (Association for Computing Machinery)¹ e IEEE (Institute of Electrical and Electronics Engineers)². Após a condução da revisão, 18 artigos foram analisados a fundo. Como o estudo de Hoef-Emden (2005) apresenta alguns conceitos teóricos de análise filogenética, ele foi analisado

¹ <https://dl.acm.org>

² <https://ieeexplore.ieee.org>

separadamente, na subseção “Conceitos Teóricos”. Por fim, os demais artigos foram categorizados com base no tipo de técnica utilizada e analisados em conjunto na subseção “Categorização”.

3.1.1 Conceitos teóricos

O estudo de Hoef-Emden (2005) contém conceitos teóricos, apresentando superficialmente alguns métodos de análise filogenética. Sendo assim, foi analisado separadamente e não foi considerado nas etapas de extração e categorização.

Hoef-Emden (2005) apresenta alguns exemplos de métodos e técnicas utilizados na literatura desde a década de 1960 para a realização do processo de análise filogenética. O primeiro estágio envolve a comparação de genes homólogos em relação ao seu comprimento. Se as sequências apresentam tamanhos diferentes, um alinhamento inicial deve ser feito.

Hoef-Emden (2005) cita métodos para geração de análises filogenéticas que exigem critérios de otimização e um algoritmo de busca em árvore. Ele destacou como critério de otimização mais comum: máxima parcimônia, matriz de distâncias e máxima verossimilhança. Esses algoritmos são geralmente combinados com algoritmos heurísticos de busca em árvore.

Por fim, Hoef-Emden (2005) apresenta desafios enfrentados pela análise filogenética utilizando grandes volumes de dados reais. Por conta dos comprimentos dos ramos altamente divergentes presentes na maioria dos conjuntos de dados, é frequente a presença de um viés na geração de árvores evolutivas. Uma forma de amenizar esse viés é adicionar sequências aos conjuntos de dados, a fim de acabar com ramos muito longos, aumentando a resolução das árvores nesses pontos.

3.1.2 Categorização

Após a leitura completa dos artigos selecionados, uma categorização foi feita de acordo com o tipo de técnica utilizada em cada um. A categorização de cada artigo está apresentada na tabela 1.

Conforme apresentado na tabela 1, cinco categorias de técnicas foram identificadas nos artigos selecionados, sendo: (I) Inteligência Artificial; (II) Ferramentas de bioinformática

existentes para alinhamento e alinhamento múltiplo; (III) Ferramentas de bioinformática existentes para geração de árvores filogenéticas; (IV) Modelos matemáticos, 3D e de cálculo de distâncias e (V) Outros tipos de técnicas.

A figura 4 apresenta a distribuição de artigos por categoria. A categoria de técnicas mais presente entre os artigos é a de ferramentas de bioinformática para alinhamento e alinhamento múltiplo; aparecendo em oito artigos. A segunda categoria de técnicas mais presente nos artigos é a de ferramentas para geração de árvores filogenéticas, com seis artigos. Técnicas baseadas em inteligência artificial e modelos matemáticos, 3D e de cálculo de distâncias foram encontradas em cinco artigos cada uma. A categoria “Outras técnicas” possui dois artigos, e optou-se por criar essa categoria pois as técnicas apresentadas nos artigos presentes nela não possuem relação entre si ou com alguma das demais categorias.

Tabela 1 – Categorização dos artigos.

Artigo	Categoria
(FANG; LUO; WANG, 2008)	(I)
(BRAMMER; WILLIAMS, 2010)	(I)
(HALGASWATHTHA et al., 2012)	(I)
(NILSSON; LARSSON; URSING, 2004)	(I) (II) (III)
(CHEN; HSIAO; TSAI, 2009)	(I) (II) (IV)
(DIXIT et al., 2008)	(II)
(PANCHAL; ANJARIA; PATEL, 2012)	(II)
(MISHRA et al., 2015)	(II) (III)
(TAN; WU, 2011)	(II) (III)
(TIAN et al., 2008)	(II) (III)
(ZHANG et al., 2009)	(II) (III)
(ZHANG; WANG, 2010)	(III) (IV)
(ZHOU et al., 2008)	(IV)
(HU et al., 2013)	(IV)
(WANG et al., 2013)	(IV)
(REQUENO et al., 2013)	(V)
(TARAWNEH et al., 2007)	(V)

Figura 4 – Distribuição de artigos por categoria.

Fonte: Priscilla Koch Wagner, 2018

Vale ressaltar que alguns artigos abordam mais de uma técnica, portanto, se encaixam em mais de uma categoria. A seguir, estão apresentados os conteúdos resumidos de cada artigo, de acordo com a sua categoria.

(I) Inteligência artificial

Dentre os artigos selecionados, cinco utilizam técnicas de inteligência artificial para análise filogenética. Desses cinco, Chen, Hsiao e Tsai (2009), Nilsson, Larsson e Ursing (2004) e Fang, Luo e Wang (2008) utilizam modelos de estados ocultos de Markov (HMM) para análise filogenética.

O trabalho de Chen, Hsiao e Tsai (2009) apresentou um método intuitivo e eficaz para extração de características e de derivação da matriz filogenética de distância, usando proteínas mitocondriais coletadas da base de dados UniProtKB 14.5³ para os experimentos. Os HMMs foram utilizados na etapa de pré-processamento das sequências, utilizando o banco de dados Pfam para recuperar os domínios funcionais das proteínas. O Pfam⁴ representa famílias de proteínas e gera agrupamentos por similaridade utilizando HMM ou alinhamentos de sequência.

No trabalho de Nilsson, Larsson e Ursing (2004), as HMMs foram utilizadas na implementação do pacote *galaxie*, que serve para identificação de sequências de forma rápida. O script *galaxieHMM* identifica e recupera sequências usando HMMs para selecionar

³ www.uniprot.org/help/uniprotkb

⁴ <http://pfam.xfam.org/>

um número arbitrário de alinhamentos. Já o trabalho de Fang, Luo e Wang (2008) utiliza a teoria de HMMs para previsão da estrutura secundária de RNA. A análise filogenética de sequências de RNA foi feita usando dois modelos de substituição evolutiva, usando teoria de HMM. O modelo para predição de estrutura é obtido integrando os modelos evolutivos (que usam HMM) com um novo perfil de gramáticas estocásticas livres de contexto.

Além de HMMs, outras técnicas de inteligência artificial foram identificadas nos artigos. É o caso do trabalho de Brammer e Williams (2010) que utiliza árvores de decisão como uma nova abordagem para estudar a convergência de uma análise filogenética. Na construção das árvores de decisão, as relações evolutivas contidas nas árvores filogenéticas são os nós internos da árvore de decisão e as folhas refletem a execução que produziu as árvores de interesse. As árvores de decisão podem medir se a construção das árvores filogenéticas está escapando de ótimos locais e garantir convergência. A profundidade da árvore de decisão determina se uma análise filogenética convergiu ou não. Quanto mais profundas, mais fortes as evidências de que a análise filogenética convergiu.

Por fim, o trabalho de Halgaswaththa et al. (2012) sugere criar uma rede neural artificial (ANN) para entender categorias de sequências desconhecidas, como dados de DNA ou proteínas, sem realizar alinhamento de sequências e gerar árvores filogenéticas. Uma rede neural probabilística e uma de retroalimentação foram utilizadas nos experimentos. A acurácia das redes neurais mostrou que as redes neurais probabilísticas podem ser métodos adequados para análise filogenética. Esse método é mais rápido que o método tradicional, pois pode compreender a categoria da sequência em apenas um passo, diferente da abordagem de análise filogenética tradicional.

(II) Ferramentas de bioinformática existentes para alinhamento e alinhamento múltiplo

Oito artigos abordaram técnicas de alinhamento ou alinhamento múltiplo. Destes, Panchal, Anjaria e Patel (2012), Nilsson, Larsson e Ursing (2004), Dixit et al. (2008), Tan e Wu (2011), Tian et al. (2008) e Zhang et al. (2009) utilizaram a ferramenta ClustalW⁵ para alinhamento múltiplo de sequências. Isso indica que esta é uma ferramenta bastante consolidada na área de análise filogenética e sua eficácia supre as necessidades dos pesquisadores.

⁵ <http://www.ebi.ac.uk/Tools/msa/clustalw2/>

Em Panchal, Anjaria e Patel (2012) a ferramenta ClustalW foi utilizada para realizar a comparação de genomas de três espécies da família *Leguminosae*. O artigo não apresentava muitas informações sobre os experimentos e resultados. O artigo de Nilsson, Larsson e Ursing (2004), já citado na categoria (I) também utiliza o algoritmo ClustalW para alinhamento múltiplo. Neste trabalho, as sequências identificadas pelo pacote *galaxie* são enviadas para o algoritmo ClustalW para alinhamento antes da geração da árvore filogenética. O objetivo de Dixit et al. (2008) em seu trabalho foi identificar proteínas relacionadas com fibrinogênio do mosquito *A. albopictus* para verificar se o mosquito possui resposta imune durante uma infecção microbiana. Para isso, realizou a extração das sequências e, em seguida, utilizou a ferramenta ClustalW para fazer o alinhamento das sequências. O trabalho de Tan e Wu (2011) propõe o PhyTree: *workbench* que integra ferramentas existentes para realizar o processo completo de análise filogenética. Dentro deste processo, a etapa de alinhamento múltiplo de sequências é realizada utilizando o ClustalW. Tian et al. (2008) realizam análise filogenética em vibrios, que são bactérias patogênicas que atacam animais criados em aquicultura. O ClustalW foi utilizado para fazer alinhamentos iniciais entre o gene *gyrB* e a sequência de 16s rRNA. Zhang et al. (2009) têm como objetivo em seu trabalho identificar as relações filogenéticas entre espécies *Laminariales*; e, para isso, utiliza o algoritmo ClustalW para fazer alinhamentos iniciais entre os genes e o 5.8s rRNA.

Mishra et al. (2015) utilizam o algoritmo Clustal Omega, parecido com o ClustalW para alinhamento múltiplo de sequências, com o objetivo de caracterizar as proteínas relacionadas a patógenos do trigo, *serpin* e *peroxidase*, avaliando-as filogeneticamente e estruturalmente.

O trabalho de Chen, Hsiao e Tsai (2009), já citado na categoria (I), apresentou um método intuitivo e eficaz para extração de características. Utiliza a ferramenta de alinhamento múltiplo da base de dados Pfam.

(III) Ferramentas de bioinformática existentes para geração de árvores filogenéticas

Seis artigos abordam ferramentas de bioinformática para geração de árvores filogenéticas. Dentre eles, Nilsson, Larsson e Ursing (2004) e Zhang e Wang (2010) utilizam a ferramenta PHYLIP para criação de árvores filogenéticas. Em (NILSSON; LARSSON;

URSING, 2004), já citado nas duas categorias anteriores, o pacote *galaxie* gera sequências que são alinhadas usando ClustalW e, em seguida, envia o resultado do alinhamento para o pacote PHYLIP realizar a análise filogenética, que resulta na exibição de uma árvore filogenética. Já o trabalho de Zhang e Wang (2010) utiliza o método de *neighbor joining* do pacote PHYLIP para geração da árvore evolutiva, com base em um conjunto de dados de 24 vertebrados.

Além da ferramenta PHYLIP, outra que foi muito citada pelos artigos foi a MEGA4.0. Mishra et al. (2015), já citados, utilizam a ferramenta MEGA4.0 para geração das árvores filogenéticas para analisar a filogenia das sequências de proteínas *serpin* e *peroxidase*. Tian et al. (2008), que também já foram citados, utilizam a ferramenta MEGA4.0 para geração de árvores filogenéticas de sequência de nucleotídeos do gene *gyrB* e de 16s rRNA, para autenticação de bactérias *vibrio*. No trabalho de Zhang et al. (2009), a ferramenta MEGA4.0 também foi utilizada para geração de árvores filogenéticas para analisar as relações filogenéticas entre espécies *Laminariales*, que é muito importante para o desenvolvimento e aplicação de recursos marinhos.

Por fim, Tan e Wu (2011) (já citados anteriormente) propõem em seu *workbench* a integração da ferramenta Tree-Puzzle, para construção de árvores filogenéticas, com as demais etapas do processo de análise filogenética.

(IV) Modelos matemáticos, 3D e de cálculo de distâncias

Modelos matemáticos, 3D e de cálculo de distância foram apresentados em cinco artigos. Chen, Hsiao e Tsai (2009), já citados anteriormente, propõem um escalonamento multidimensional com base na matriz de distâncias, gerando uma visualização em espaço 3-D das coordenadas relativas das espécies, pois árvores filogenéticas podem ser menos informativas em termos de dimensionalidade. O trabalho organiza o conjunto de dados na forma de um cubo (espécies x proteínas x domínios funcionais).

O trabalho de Zhang e Wang (2010) usa a distância *Bhattacharyya*, uma medida de distância teórica entre duas distribuições de probabilidade, como base de classificação para analisar a filogenia de proteínas de 24 vertebrados.

Zhou et al. (2008) apresentam uma análise filogenética baseada em rede. Nesses casos, a distância evolutiva pode ser definida como a diferença entre os conjuntos de

elementos da rede. Dessa forma, o artigo propõe uma nova abordagem para determinar a distância evolutiva e obter quatro novas medidas de distância, sendo comparadas com a distância de *Jaccard*. Os resultados mostraram que as novas medidas de distância cometem erros menores em todos os casos de teste de comparação; indicando que a distância evolutiva é decidida também pela conservação filogenética e importância topológica.

No trabalho de Hu et al. (2013) foi proposto um método matemático para caracterizar uma sequência de DNA como um vetor de características. A distância entre dois genes é obtida através do cálculo da distância entre os seus vetores de características. Isso gera um espaço vetorial com a distância biológica, o que permite determinar a distância evolutiva entre espécies e fazer análise evolutiva de maneira mais fácil. Este método é livre de alinhamento e, por isso, mais rápido que os métodos à base de alinhamento convencional.

Por fim, com o objetivo de realizar análises filogenéticas usando sequências genômicas integrais, Wang et al. (2013) propõem um método de computação de distância (método JCV) para análise filogenética de sequência de DNA. O método não requer alinhamento múltiplo de sequências, por isso, é capaz de calcular a distância entre sequências em todo nível do genoma.

(V) Outros tipos de técnicas

Por último, nesta categoria se encontram os dois trabalhos que não se encaixaram nas categorias anteriores e tampouco apresentam relação entre si.

O trabalho de Requeno et al. (2013) utiliza verificação de modelos (*Model Checking*) para comparação de árvores filogenéticas com as especificações de suas propriedades biológicas. O modelo verifica automaticamente a corretude do sistema da árvore filogenética.

Tarawneh et al. (2007) apresentam em seu trabalho uma análise filogenética do táxon *excavates* e seus familiares, usando o método já desenvolvido QBNJ. Esse método é um processo de aglomeração para construção de árvores binárias a partir de um conjunto de sequências moleculares. Experimentos que utilizam um grande volume de dados mostram que QBNJ pode superar muitos métodos existentes.

3.1.3 Considerações finais da revisão sistemática

Os resultados obtidos mostraram que existem algumas ferramentas tradicionais consolidadas que são amplamente utilizadas na análise filogenética. Alguns artigos encontrados utilizaram tais ferramentas: (PANCHAL; ANJARIA; PATEL, 2012), (NILSSON; LARSSON; URSING, 2004), (ZHANG; WANG, 2010), (DIXIT et al., 2008), (MISHRA et al., 2015), (TAN; WU, 2011), (TIAN et al., 2008) e (ZHANG et al., 2009). Por conta disso, tais artigos não apresentam avaliação do método utilizado, assumindo que os resultados fornecidos por essas ferramentas já estão aprovados pela comunidade científica da área. Apesar de existirem ferramentas consolidadas na área de análise filogenética, os esforços para sua melhoria estão em falta.

Verificou-se que algumas iniciativas propõem diferentes abordagens para análise filogenética, tais como ferramentas de inteligência artificial e modelos matemáticos, que podem substituir ou complementar esses métodos tradicionais. No entanto, tais abordagens ainda são pouco expressivas. Isso pode ser destacado nos estudos de: Fang, Luo e Wang (2008), que usa teoria de HMMs e gramáticas estocásticas livres de contexto para prever estruturas de rRNA secundárias; Brammer e Williams (2010), que usa árvores de decisão para análise filogenética; Halgaswaththa et al. (2012), que se propõe a analisar categorias de sequências desconhecidas usando redes neurais artificiais; Chen, Hsiao e Tsai (2009), que propõe a visualização de coordenadas relativas a espécies, proteínas e domínios funcionais, no espaço 3D; Zhou et al. (2008), que propõe uma nova abordagem para determinar distâncias evolucionárias; Hu et al. (2013), que propõe um método matemático para definir sequências de DNA como um vetor de características; Wang et al. (2013), que propõe um método de computação de distância que não requer alinhamento múltiplo de sequências. As avaliações dos métodos propostos nesses artigos, em geral, eram mais rigorosas do que as apresentadas nos artigos que utilizam as ferramentas tradicionais.

Poucas abordagens propõem melhorias e integração automática das ferramentas que implementam etapas do processo de análise filogenética. Atualmente, a comunicação entre essas etapas é feita manualmente, o que torna o processo muito custoso para os pesquisadores. Neste ponto, vale ressaltar o trabalho de Tan e Wu (2011) que propõe o Phytree, um workbench que integra Blast, ClustalW e Tree-Puzzle, para pesquisa de

sequências homólogas, alinhamento múltiplo e reconstrução da árvore filogenética. Apesar disso, a avaliação da abordagem do Phytree no trabalho de Tan e Wu (2011) foi superficial.

Esta revisão forneceu uma visão geral do estado da arte para análise filogenética e apresenta detalhes e metodologias para a sua reprodução e atualização. As análises feitas durante a revisão confirmaram a carência de uma ferramenta integrada para análise filogenética.

3.2 Análises considerando genes exclusivos

O presente trabalho possui foco em genes considerados exclusivos em análises de genomas próximos filogeneticamente. Por conta disso, sentiu-se a necessidade de identificar trabalhos sobre genes exclusivos para embasar o desenvolvimento deste projeto. Nesta seção, serão apresentados alguns trabalhos encontrados na literatura sobre genes exclusivos.

Moreira et al. (2010) comparou o conteúdo genômico de três cepas que causam tipos diferentes de câncer em espécies diferentes de vegetais cítricos: *Xanthomonas citri* subsp. *citri* (causa câncer do tipo A), *Xanthomonas fuscans* subsp. *aurantifolii* strain B (causa câncer do tipo B) e *Xanthomonas fuscans* subsp. *aurantifolii* strain C (causa câncer do tipo C). A partir dessa comparação, foram identificados genes comuns às três cepas: *pthA*, *xopE3* e *xopAI*. Tais genes são bons candidatos para serem fatores de patogenicidade em cítricos. Além disso, existe muitas diferenças de genes entre as três cepas, que podem ser correlacionadas com sua diferente virulência e vegetais hospedeiros. Os resultados obtidos nesse estudo podem ser empregados para determinar corretamente o papel desses genes exclusivos de cada cepa nas interações entre planta e bactéria patogênica, a fim de melhorar o controle do câncer cítrico.

Como já comentado, o estudo de genes exclusivos pode ser utilizado na identificação do uso correto de antibióticos ou vacinas e para a realização do tratamento mais adequado. Nessa linha, existe o trabalho de Nogueira et al. (2017), onde foram comparadas cepas de *Mycobacterium chelonae*-*M. abscessus* complex (do cluster CHII), utilizando diversas abordagens, analisando fenotipicamente, por teste de susceptibilidade a fármacos, sequenciamento do genoma completo e análise de um único gene. A variabilidade observada durante o estudo sugeriu a presença de diferentes grupos taxonômicos entre os genomas isolados.

As análises fenotípicas e os testes de suscetibilidade a fármacos não foram informativos para distinguir os grupos taxonômicos identificados pelas análises genômicas.

Não foram encontrados muitos estudos que analisam especificamente genes exclusivos de espécies, apesar de muitos estudos compararem espécies próximas para identificar especificidades.

3.3 Análises considerando bactérias do gênero *Xanthomonas*

Como apresentado no capítulo 2, bactérias do gênero *Xanthomonas* vem sendo amplamente estudadas pela comunidade científica, visto que o melhor entendimento dessas bactérias pode trazer incontáveis benefícios para a agricultura, já que pode ajudar a tratar e prevenir doenças nas espécies vegetais que ela ataca. Esta seção apresenta alguns estudos relacionados a este trabalho, que utilizam bactérias do gênero *Xanthomonas*.

Primeiramente, a fim de ilustrar a importância dos estudos das espécies do gênero *Xanthomonas*, Ryan et al. (2011) apresenta em seu artigo uma revisão de trabalhos genômicos funcionais e comparativos da adaptação desse grupo de bactérias para explorar a diversidade de plantas hospedeiras. As espécies de *Xanthomonas* apresentam um alto grau de especificidade nas plantas hospedeiras e seus tecidos, difíceis de serem identificados. O poder da comparação de genomas para definir a causa das especificidades do hospedeiro e dos tecidos, bem como os genes essenciais para a patogênese, é proporcional ao número de sequências do genoma disponíveis. Portanto, o sequenciamento dos genomas de uma maior diversidade de cepas e espécies do gênero *Xanthomonas* é de extrema importância para os estudos dessas bactérias, a fim de desenvolver tratamentos para as doenças causadas por elas.

O estudo de Jalan et al. (2013) comparou os genomas de duas sub-espécies de *Xanthomonas*: *Xanthomonas citri* subsp. *citri* strain 306 (XccA306) e *Xanthomonas citri* subsp. *citri* strain A(w)12879 (Xcaw12879) a fim de encontrar as especificidades de cada uma. Como resultado, foram identificados genes que influenciavam na limitação de hospedeiros da sub-espécie Xcaw12879 e outros que contribuem para a virulência; além de diversos genes que possuem expressões diferentes para ambas espécies.

O trabalho de Moreira et al. (2010), já citado na seção 3.2 deste capítulo, estuda o cancro cítrico em seu trabalho. Nele, três bactérias dessa sub-espécie foram estudadas

pois cada uma causa um tipo de cancro diferente em plantas cítricas: *Xanthomonas citri* subsp. *citri* (previamente sequenciado), *Xanthomonas fuscans* subsp. *aurantifolii* strain B e *Xanthomonas fuscans* subsp. *aurantifolii* strain C. As espécies B e C foram sequenciadas e comparadas com o genoma da espécie já sequenciada com o objetivo de identificar genes candidatos para ser fator de patogenicidade em cítricos e identificar as diferenças entre os genes existentes nas três cepas.

Diversos estudos visam a identificar genes resistentes a doenças ou formas de tratar patologias. O estudo de Xia et al. (2010), por exemplo, utiliza técnicas de Máquinas de Vetores de Suporte (*Support Vector Machine* - SVM) para prever a chance de um proteína do arroz ser resistente a *Xanthomonas oryzae* pv. *oryzae*, a partir de dados das sequências de genes do vegetal e da bactéria.

Outro trabalho relacionado é o apresentado por Borges et al. (2016), que visa a detectar e classificar a gravidade de doenças em campos de tomate. Nele, não foram utilizados dados genéticos para estudo das *Xanthomonas*. Por sua vez, foram utilizadas imagens de subáreas de plantações de tomates e, a partir dessas imagens, foram utilizadas técnicas de processamento de imagens, segmentação e agrupamento para identificar e classificar manchas bacterianas causadas pelas espécies *Xanthomonas perforans* e *Xanthomonas gardneri*. Os resultados apresentaram forte correlação com as análises de especialistas e menor erro ao analisar quanto a doença afeta a produtividade, além de possuir a vantagem de possuir baixa variação, alta velocidade e fácil aplicabilidade em campo.

O estudo de Zhang et al. (2015) estuda a história evolutiva da espécie *Xanthomonas axonopodis* subsp. *citri*, que ataca espécies cítricas. Nesse trabalho, foram sequenciados 21 indivíduos e comparados genomicamente e evolutivamente. As análises sugerem que a aquisição de genes benéficos e a perda de genes prejudiciais ao longo do tempo permitiu que essa sub-espécie de *Xanthomonas* passasse a infectar uma gama maior de hospedeiros. Os resultados desse estudo colaboram no tratamento e prevenção de novas doenças.

Problemas relacionados a alinhamento de genomas são utilizados para avaliar performance de *hardware*, por conta da complexidade do problema. O estudo de Almeida et al. (2003), por exemplo, utilizou computação paralela de alta performance para comparar genomas completos de duas bactérias do gênero *Xanthomonas*, *Xanthomonas axonopodis* pv. *citri* e *Xanthomonas campestris* pv. *campestris*, ambas com mais de 5 milhões de pares de bases. O estudo teve dois objetivos principais: analisar o poder de computadores de baixo custo em uma solução paralela para esse problema; comparar também regiões e

lacunas entre os genomas, além dos genes homólogos. A solução proposta apresentou menor tempo de processamento, além de resultados com maior qualidade para a comparação desses genomas, em comparação com Blast.

4 Abordagem para Análise Filogenética de Genes Exclusivos

Neste trabalho, o problema de identificação da origem de genes exclusivos foi tratado a partir do desenvolvimento de uma abordagem de análise proposta. A origem de genes exclusivos é uma questão importante na área de biologia comparativa, pois, a partir de análises de um gene (ou número pequeno de genes), é possível identificar espécies ou cepas próximas geneticamente, podendo auxiliar, por exemplo, no processo de identificação do tratamento mais indicado para uma doença ou na procura por alvos moleculares para a criação de vacinas (NOGUEIRA et al., 2017).

Durante o desenvolvimento inicial deste trabalho, e como forma de prova de conceito, foi desenvolvido um estudo de caso, cujos resultados estão apresentados na seção 5.2 do capítulo 5. Este estudo de caso visou a analisar a taxonomia de uma espécie de *Xanthomonas*, a *Xanthomonas campestris* pv. *campestris*.

Os genes exclusivos da bactéria *Xanthomonas campestris* pv. *campestris* foram obtidos a partir de uma ferramenta desenvolvida pelo grupo de pesquisa (PEREIRA, 2017), que realiza a comparação de espécies, a partir de um arquivo de alinhamentos de sequências (neste projeto o programa de alinhamentos locais BLAST (ALTSCHUL et al., 1990) foi utilizado). Essa ferramenta compara os genomas a fim de identificar os genes homólogos, sendo possível também identificar famílias específicas de genes como, por exemplo, genes exclusivos a um genoma. Tal ferramenta compara todos os genomas contra todos os genomas, sendo que o gene irá pertencer a uma família de genes se ele possuir um alinhamento (respeitando-se alguns limiares) com algum gene pertencente a essa família em outro genoma que não o seu original. Dessa forma, os genes exclusivos de um genoma são identificados como os que não foram alinhados com nenhuma família de genes dos demais genomas.

A seguir, esses genes exclusivos foram utilizados como genes de referência para comparação e alinhamento com os genomas do banco do NCBI, também utilizando a ferramenta BLAST. Após esse passo, um arquivo no formato M8 (explicado na seção 4.1 deste capítulo) foi obtido, contendo as informações dos alinhamentos entre os genes exclusivos da *Xanthomonas campestris* pv. *campestris* e os genes “semelhantes” dos demais genomas do banco do NCBI.

Nesse ponto, foi realizada uma análise manual do resultado dos alinhamentos e foi possível perceber que os genes de referência alinharam (considerando alguns limiares de qualidade dos alinhamentos) apenas com genes da própria espécie *Xanthomonas campestris* pv. *campestris*. Ou seja, os genes exclusivos da bactéria *Xanthomonas campestris* pv. *campestris* não estão presentes em nenhum outro genoma do banco do NCBI.

Mesmo sendo genes exclusivos, quando analisado um grupo de espécies próximas, é incomum que genes exclusivos de uma espécie não possuam nenhum alinhamento relevante com outros genomas. Isto posto, algumas hipóteses foram levantadas para explicar tal comportamento, para serem analisadas no decorrer deste trabalho:

- os genes exclusivos derivaram de outras famílias de genes do próprio organismo;
- os genes se diferenciaram muito dos genes ancestrais e por isso não se alinham com outras espécies;
- há uma inconsistência na anotação dos genes.

Como já comentado, a propagação de erros de anotação é um dos grandes problemas enfrentados por pesquisadores de biologia comparativa, dado que pode afetar outros processos que se utilizam direta e indiretamente dos resultados dessa etapa (MOREIRA, 2015). Por esse motivo, decidiu-se por analisar primeiramente a terceira hipótese, a fim de remover as possíveis inconsistências nas anotações dos genes utilizados durante todo o desenvolvimento do projeto. As outras duas hipóteses também foram analisadas após a análise das anotações.

Conforme apresentado na figura 5, o processo de análise de genes exclusivos desenvolvido nesse projeto se inicia a partir de arquivos com todos os genes anotados e com todas as sequências de nucleotídeos dos genomas sob análise. Esses arquivos são comparados, utilizando o BLAST para realização dos alinhamentos (genes alinhados com genes de genomas próximos e sequências de nucleotídeos alinhadas com sequências de nucleotídeos de genomas próximos).

Os arquivos com os alinhamentos resultantes foram utilizados para a identificação dos potenciais genes exclusivos, resultando em uma matriz de presença de genes e uma matriz de presença sequências de nucleotídeos. Com essas duas matrizes foi realizada a análise de anotação dos genes considerados exclusivos, resultando em uma lista com os genes realmente exclusivos de cada genoma em análise.

Figura 5 – Diagrama completo da abordagem desenvolvida nesse projeto.

Fonte: Priscilla Koch Wagner, 2018

A seguir, foi realizada a comparação entre esses genes exclusivos e as sequências de nucleotídeos de cada genoma do estudo, utilizando BLAST. Por fim, a partir do arquivo de alinhamento, foi realizada uma análise de famílias de genes no próprio genoma onde cada gene exclusivo foi encontrado; além de uma análise da taxonomia de cada espécie em relação ao banco do NCBI.

Todas as ferramentas desenvolvidas durante este trabalho foram implementadas na linguagem Java. Alguns scripts em Python e em R foram desenvolvidos em paralelo para tratamento de arquivos de dados e visualização de dados. Além disso, como sabido, ferramentas computacionais padrão da área de biologia comparativa e genética foram utilizadas, como o BLAST.

Este capítulo apresenta em detalhes as etapas desenvolvidas durante esse projeto para a análise de genes exclusivos. Primeiramente, a seção 4.1 apresenta uma descrição das representações de dados utilizadas no decorrer do desenvolvimento desse trabalho. A seguir, a seção 4.2 descreve a etapa preliminar de identificação dos potenciais genes exclusivos dos genomas; a seção 4.3 apresenta a etapa de identificação dos erros de anotação em genes considerados exclusivos e confirmação dos genes realmente exclusivos; a seção 4.4 apresenta a identificação de famílias de genes similares aos exclusivos dentro do próprio genoma e a seção 4.5 apresenta a implementação da identificação da taxonomia relacionada aos genes exclusivos em relação aos genomas encontrados no banco do NCBI.

4.1 Representação dos dados

Durante a implementação da abordagem de análise filogenética de genes exclusivos desenvolvida neste projeto de mestrado, foram utilizados ou desenvolvidos alguns arquivos com informações relacionadas aos genes, genomas e sequências analisadas no processo. Tais arquivos e suas respectivas informações são descritos nesta seção.

4.1.1 Arquivo de alinhamento - M8

Conforme apresentado no capítulo 2, o BLAST (ALTSCHUL et al., 1990) é uma ferramenta de alinhamento local, muito utilizada por pesquisadores de filogenética, biologia comparativa e evolutiva, que busca regiões similares em sequências de genomas. Esta

ferramenta foi utilizada ao longo deste projeto, para realizar comparações entre genes, genomas e sequências de nucleotídeos.

O resultado do processamento dos alinhamentos no BLAST é um arquivo no qual cada registro possui informações a respeito de um alinhamento. Neste projeto utilizou-se o formato de arquivo M8 que, no total, possui 12 campos com informações de cada alinhamento, descritos a seguir:

- *query*: identificador da sequência fornecida como entrada;
- *subject*: identificador da sequência da base de dados que foi alinhada com a sequência de entrada;
- *percent identities*: porcentagem da identidade do alinhamento, ou seja, da ocorrência, ao longo do comprimento do alinhamento, de mesmos nucleotídeos ou aminoácidos nas mesmas posições no alinhamento das duas sequências;
- *alignment length*: comprimento do alinhamento;
- *mismatches*: número de posições em que houve alinhamento de nucleotídeos ou aminoácidos diferentes;
- *gap openings*: número de lacunas (sequências de espaços em branco) inseridas no alinhamento;
- *query sequence start*: posição inicial do alinhamento na sequência de entrada;
- *query sequence end*: posição final do alinhamento na sequência de entrada;
- *subject sequence start*: posição inicial do alinhamento na sequência encontrada na base de dados;
- *subject sequence end*: posição final do alinhamento na sequência encontrada na base de dados;
- *e-value*: probabilidade do alinhamento ter ocorrido por coincidência. Quanto menor o e-value, menores são as chances do alinhamento ter ocorrido pelo acaso;
- *bit-score*: nota normalizada derivada do score. É a nota dada ao alinhamento, baseada no pareamento de nucleotídeos/aminoácidos idênticos e diferentes entre as sequências.

A partir desses campos é possível fazer diversas análises a respeito dos alinhamentos. Por exemplo, definir a cobertura do alinhamento, ou seja, quantas bases foram alinhadas em relação ao tamanho total da sequência original (independente da posição das bases). Isso pode ser calculado a partir do *query sequence start* e *query sequence end* das sequências que alinharam com a sequência original e o *alignment length* da sequência original.

4.1.2 Matriz de Presença

Como citado anteriormente, uma das etapas deste projeto foi identificar os genes exclusivos dos genomas estudados. Para isso, foram utilizadas, inicialmente, matrizes de presença que relacionam todos os genes presentes nos genomas estudados, ou seja, matriz *gene vs. genoma*.

Essa matriz é configurada de forma que: a primeira linha contém os genomas utilizados durante as análises, enquanto a primeira coluna contém todos os genes presentes nesses genomas. Cada célula interna da matriz apresenta a quantidade de vezes que o gene indicado na linha está presente no genoma indicado pela coluna. Caso um gene apresente o valor 1 em uma coluna, e 0 nas demais, ele é considerado exclusivo. A figura 6 abaixo exemplifica uma matriz de presença.

Figura 6 – Exemplo da configuração de uma matriz de presença.

gene	genoma1	genoma2	genoma3	genoma4	genoma5
gene1	1	1	0	0	1
gene2	0	1	0	1	0
gene3	0	0	0	0	1
gene4	0	0	0	3	0
gene5	0	0	1	0	0

Fonte: Priscilla Koch Wagner, 2018

Na figura 6, os genes ‘gene1’ e ‘gene2’ não são exclusivos pois estão presentes em mais de um genoma; o ‘gene4’ não é exclusivo pois está presente mais de uma vez no mesmo genoma. Já os ‘gene3’ e ‘gene5’ são exclusivos pois estão presentes em apenas um genoma.

Outra matriz de presença, com a mesma configuração descrita acima, foi gerada a partir das sequências de nucleotídeos. Nessa matriz, a primeira linha também contém os genomas utilizados nas análises e a primeira coluna, os genes presentes nesses genomas. Porém, as células internas indicam a quantidade de vezes que qualquer sequência de nucleotídeos (não apenas as sequências anotadas como gene), similar ao gene indicado na linha, se apresenta em cada genoma. Essa matriz foi chamada de *nucleotídeos vs. genoma*.

4.1.3 FASTA

No processo de análise dos erros de anotação encontrados, foram utilizados arquivos no formato FASTA. Esses arquivos, com a extensão .FNA (quando se trata de sequências de nucleotídeos) ou .FAA (quando se trata de sequências de aminoácidos), contém informações a respeito das sequências de DNA, que podem ser utilizadas em *softwares* de biologia molecular.

Um arquivo FASTA pode ser composto por muitas sequências de nucleotídeos. Cada sequência no formato FASTA começa com uma linha de descrição, com informações sobre o genoma de referência da sequência, tamanho da sequência. Após a linha de descrição, seguem algumas linhas com os dados da sequência de nucleotídeos em si. A identificação de uma nova sequência de nucleotídeos no arquivo FASTA é feita a partir do caracter '>', que inicia a linha de descrição da sequência.

A figura 7 apresenta trechos de arquivos FASTA utilizados no presente projeto. A imagem (a) da figura 7 é um trecho de um arquivo no formato .FNA, enquanto a imagem (b) é um trecho de um arquivo no formato .FAA.

4.2 Identificação dos Potenciais Genes Exclusivos

A abordagem desenvolvida neste projeto de mestrado visa a análise da filogenia de genes exclusivos. Para isso, o primeiro passo desenvolvido foi a identificação dos potenciais genes exclusivos, a partir da análise de alinhamentos dos genes de genomas de espécies próximas. Esta seção visa a descrever este processo.

A ferramenta que implementa esse processo foi desenvolvida pelo grupo de pesquisa em que este projeto se insere e foi utilizada durante o desenvolvimento deste trabalho. Essa ferramenta objetiva a comparação de espécies; portanto, tem como resultado, além da lista de genes exclusivos e matriz de presença de genes, utilizados ao longo deste trabalho, outros resultados comparativos entre as espécies em estudo. Além disso, ela foi desenvolvida visando a análise de alinhamentos entre genes, mas, nesse projeto também foi utilizada para geração da matriz de presença de sequências de nucleotídeos, além da matriz de presença de genes.

Figura 7 – Trechos de arquivos FASTA. (a) trecho no formato .FNA. (b) trecho no formato .FAA.

```
>gi|471265562|ref|NC_020815.1|:42-1370 Xanthomonas citri subsp. citri Aw12879, complete genome
ATGGATGCTTGGCCCCGCTGTCTGGAACGCTCTGAAGCCGAATTCGCCCGGAAGATGTCCACACCTGGT
TGAACCCCTGCAAGCCGAAGATCGCGGCGACAGCATGTGTGTACGCTCCGAACGCTTCATCGTCGA
CCAGGTGCGCGAACGTTACCTGCCGCGCATTCGCGAATTGCTGGCGTATTTGCTCGGCAATGGCGACGTC
GCGCTGGCGGTGCGTTGCGGTCCACGCGCGCCGAGCCGGCCCTGCACCCGTGGCCGTGCCGAGCGCGC
CGCAGGTGCAACGATCGTGCCGTTCCGCCGCAACCTGGATTGCGATTACACCTTCGCCAAGTTCGTGGA
AGGCCGCAAGAACGTTGGGCTGGCGGCGCGATCCAGGTGCGACAGAAGCCCGGCGACCGTGACAC
AACCCGTTGCTGTGTACGGCAGCACCAGTCTGGGCAAGACCCACCTGATGTTGCGGCGCGCAATGCGC
TGCCTCAGGCCAATCCGGCCGCCAAGGTGATGTATCTGCGTTGCGGAGCAGTTCTTCAGCGCGATGATCCG
GGCGTGCAGGACAAGGCGATGGACAGTTCAAGCGCCAGTTCCAGCAGATCGATGCGCTGCTGATCGAC
GACATCCAGTTCTTCGCCGCGCAAGGACCGACCCAGGAAGAGTTCTCCACACCTTCAACGCGCTGTTTCG
ATGGCGCGCAACAGATCATCTGACCTGCGACCGCTACCCGCGCGAGGTGGAAGGCTGGAGCCGCGGCT
GAAGTGCAGCTGGCTGGGCTGTCGGTGGCGATCGACCCGCCGACTTCGAGACCCGCGCGGCCATC
GTGCTGGCCAAGGCGCGCGAGCGCGCGCGGCGAAATTCGCGACGAGTGGCATTCTGATCGCCAAAGAAGA
TGCGCTCCAACGTGCGCGACCTTGAAGGCGCGCTCAACACGCTGGTGGCGCGTGCCAACTTCAACGGGCG
CTCGATCACGGTGGAAATTTGCCAGGAGACGCTGCGCGACCTGCTGCGCGCCAGCAACAGGCGATCGGC
ATTCCCAATATCCAGAAGACTGTGGCCGATTACTACGGCTTGCAAATGAAGGATCTGCTGTCCAAGCGGC
GCACCCGCTCGCTGGCGCGCCGCCAGGTGGCGATGGCGCTGGCCAAGGAACCTACCGAGCACAGCCT
GCCGGAGATCGCGATGCTTCGCTGGCGCGACACACCCAGTGTGATGCGTGCCGTGAGATCCGC
ACGTTGATGGAGGCGGACGGCAAGCTCCGCGAGGACTGGGAAAAGCTGATTCGCAAGCTCAGCGAGTAA
```

(a)

```
>gi|471265593|ref|YP_007648058.1| Cellulase [Xanthomonas citri subsp. citri Aw12879]
MSAVCFSLRALLRSTLLFIVFGSLPLAHAQTRALTYAGVNLSGAEFASSKKPGVLNKDYMMAASDYTYF
AGVGMINTIRLPILWERLQPTASGELDPAQLALLQAVARAKAAGMYLVIDIHNYAKYYGYKIGSPEVPVA
TFTDLWRRLALAFNSDNAVMFGLMNEPNNISASDWAGAAQAIDAIRRTGANNLILVPGALWTGAHSWYS
TTNDGYSNATALTSIYDPLDRYAFEVHQYLDADSSGTSSTCVSSTIGAERLRNFTQWLRNTRKRGFLAEF
GTANNAVCNQALQGLMAYMETNADVWLGNWAAAGSWMNL SYAYSVHPNKDGTDPKQMAILSPQATLATR
APVAAAAARTSAASTSATSTSVARTAR
>gi|471265594|ref|YP_007648059.1| Cellulase [Xanthomonas citri subsp. citri Aw12879]
MKSLTTHTRRLRCRALSFLLLAAMPLAQASATGLKYAGVNLSGAEIKSSQKPGVLNIDYRYPAADEYAY
FAGKHMINTVRLPILWERLQPRAGGELDPAQLALIRQAVANAKAAKMYVIDVHNYAKYYGHTIGSTKVPI
STFNDLWRRLAIAFKSDNAVIFGLMNEPYDIDPQAWATAAQASIDTIRKTGATNLILVPGALWTGAHSWY
STVAGQSNVAMASIRDPLNRYAIEAHQYLDADSSGTSGGCVSATVGVERLSFTEWLRNLNKKRGFLAEF
GTGNTVTCLNALNGMLGYMESNSDVWIMGHSWAAAGAWNSGAYPFNVQPDAGROKPMQMSILSPARRITD
```

(b)

Fonte: Priscilla Koch Wagner, 2018

Primeiramente, foram realizados alinhamentos, utilizando o BLAST, entre todos os genes de todos os genomas envolvidos na análise (todos contra todos). A partir da leitura desses alinhamentos, a ferramenta seleciona os melhores alinhamentos de cada gene de todos os genomas, a partir de alguns filtros e marca os genes alinhados como semelhantes. Por exemplo, se o gene1 (do genomaX) possui três alinhamentos relevantes, com o gene2 (do genomaY), o gene3 (do genomaZ) e o gene4 (do próprio genomaX), os genes 2, 3 e 4 são marcados como semelhantes do gene1 em seus respectivos genomas de origem.

A seguir, para cada gene de cada organismo identificado pela ferramenta, são feitas análises, em comparação com seus semelhantes. Para geração da matriz de presença, é feita a contagem de cada gene, no próprio genoma de origem e nos genomas em que aparecem genes semelhantes. Por fim, os genes analisados são marcados, para evitar que

Figura 8 – Comparação de genomas e geração da matriz de presença com contagem de genes por genoma.

Fonte: Priscilla Koch Wagner, 2018

os genes semelhantes sejam contados novamente quando o genoma de origem estiver sendo analisado.

Usando novamente o exemplo citado no parágrafo anterior: ao varrer os genes do genomaX, encontra-se o gene1 e soma-se 1 na contagem desse gene no genomaX. A seguir, os genes marcados como semelhantes a ele são analisados. Soma-se 1 na contagem do gene1 no genomaY e 1 na contagem do gene1 no genomaZ. Como o gene1 possui um semelhante no próprio genomaX (gene4), é somado mais 1 à contagem desse gene nesse genoma. A figura 8 apresenta o fluxo implementado para geração da matriz de presença de genes, aplicada ao exemplo descrito.

A identificação dos potenciais genes exclusivos nesse processo é feito durante a leitura do arquivo de alinhamento. Os genes que não possuem nenhum alinhamento válido,

além do seu próprio, podem ser considerados exclusivos. Isso é confirmado no momento da contagem dos genes: um gene que possui apenas ele mesmo como semelhante, será contado apenas uma vez no seu próprio genoma de origem.

4.3 Identificação de erros de anotação e genes realmente exclusivos

Conforme apresentado no capítulo 2, o processo de anotação genética é passível de erros, fazendo com que parte das sequências em bancos de dados públicos de genomas, que servem de base para estudos de genômica, apresente inconsistências nas anotações.

Partindo da hipótese de que os genes exclusivos podem ter sido anotados erroneamente nos genomas das bactérias do gênero *Xanthomonas*, faz-se necessária uma etapa de verificação das anotações, antes da análise da origem dos genes das bactérias de interesse em relação aos genomas do banco do NCBI.

A verificação da anotação genética foi realizada a partir de duas etapas, implementadas em duas ferramentas, sendo que ambas possuem resultados complementares e fazem parte de processos distintos dentro dessa abordagem. Esta etapa está apresentada na figura 9. Esta etapa contempla estudos para identificação de genes exclusivos de maneira genérica bem como auxilia na identificação de erros de anotação (exclusivos ou não).

A primeira tarefa da etapa de análise dos genes exclusivos (representada pelo fluxo (a) da figura 9) visa a analisar os potenciais genes exclusivos e identificar possíveis erros de anotação. Além de identificar os erros de anotação, essa tarefa segmenta os dados a fim de identificar se o erro de anotação é relativo a outras sequências dentro do próprio genoma (erros intragenoma) ou encontrados em outros genomas (erros intergenoma).

A ferramenta que implementa essa tarefa consiste na análise de duas matrizes de presença (descritas na seção 4.1 deste capítulo) como entrada. A primeira matriz contém, para cada gene anotado de cada genoma a ser estudado, o número de vezes em que foram encontrados genes semelhantes entre os genes anotados dos genomas das espécies filogeneticamente próximas sendo estudadas. A segunda matriz apresenta a mesma configuração, porém, contém, para cada gene anotado, o número de sequências de nucleotídeos semelhantes encontradas nos cromossomos dos genomas de espécies filogeneticamente próximas.

Figura 9 – Análise de inconsistências no processo de anotação genética.

Fonte: Priscilla Koch Wagner, 2018

A ferramenta efetua a leitura das duas matrizes, e compara as presenças dos genes com as presenças das sequências de nucleotídeos nos genomas a fim de encontrar inconsistências, visto que a premissa é de que se um gene for efetivamente exclusivo, não se deve encontrar uma sequência de nucleotídeos similar a esse gene nos cromossomos (exceto a própria sequência que sintetiza esse gene no genoma em que ele é exclusivo). As inconsistências correspondem a uma sequência de nucleotídeos, que havia sido considerada gene exclusivo em um genoma, estar presente em mais de um genoma.

Se o gene foi anotado em apenas um genoma e, também, não estiver presente como sequência de nucleotídeos em outros genomas (ou seja, não possuir uma sequência de nucleotídeos não anotada em outros genomas), esse gene é considerado exclusivo. Caso

contrário, é identificado como possível erro de anotação a ser analisado e considerado não-exclusivo.

Como resultado, essa ferramenta retorna quatro arquivos, em formato .csv. O primeiro apresenta uma listagem dos genes que não apresentaram nenhum erro de anotação, e portanto continuam sendo considerados genes exclusivos. O segundo arquivo consiste em uma listagem dos genes que apresentaram potenciais erros intergenomas (ou seja, que possuíam uma sequência semelhante em outros genomas), e a quantidade de vezes que esse erro potencial foi identificado (em quantos genomas existem sequências similares). O terceiro arquivo também consiste em uma lista das genes com potenciais erros de anotação, porém contendo a quantidade de vezes que a sequência foi identificada no próprio genoma, sendo chamados de potenciais erros intragenomas. Por fim, o quarto arquivo apresenta a distribuição de erros de anotação, sejam intergenomas ou intragenomas, pelos genomas em estudo.

A segunda tarefa da etapa de análise dos genes exclusivos (representada pelo fluxo (b) da figura 9) atua de maneira a complementar os resultados da primeira. Realiza uma dupla checagem a fim de identificar quais das sequências de nucleotídeos que se alinharam aos genes anotados efetivamente não deveriam ser anotados como gene. Isto é, dados os potenciais erros de anotação identificados, esta ferramenta visa a verificar se efetivamente a sequência de nucleotídeos poderia sintetizar uma proteína (ou seja, o gene em si não era exclusivo e a sequência de nucleotídeo comparada deveria ser anotada como gene) ou se é apenas uma sequência parecida com o gene anotado (ou seja, o gene anotado é realmente exclusivo).

A ferramenta que implementa essa tarefa utiliza um arquivo de alinhamentos do BLAST (ALTSCHUL et al., 1990), com os alinhamentos dos genes sob análise com os genomas filogeneticamente mais próximos da espécie sendo analisada. Além do arquivo de alinhamentos, também são utilizados arquivos FASTA, que possuem as sequências de nucleotídeos dos genomas de referência que foram utilizados nos alinhamentos.

Para cada gene foi identificada qual a sequência que melhor se alinhou, exceto o próprio gene. A seguir, verificou-se se a sequência atende aos parâmetros de identidade e cobertura definidos no momento da execução. Caso a sequência atenda aos parâmetros definidos, ela é analisada em relação à identificação de códons de parada.

A identificação de códons de parada objetiva garantir que a sequência de nucleotídeos sob estudo pode ser um gene, pois a sequência de nucleotídeos que codifica aminoácidos

para sintetizarem proteínas não podem possuir um códon de parada em seu interior. É necessário verificar se existe, dentro dessa sequência de nucleotídeos, uma subsequência de bases nitrogenadas que seja considerada um códon de parada. Para tal, é realizada a leitura do gene em três partes: sempre seguindo de três em três bases nitrogenadas, a partir da primeira base, a partir da segunda base e a partir da terceira base. E então repetindo o mesmo processo para o seu complemento reverso, totalizando assim em seis possibilidades para que um códon de parada seja encontrado na sequência.

Para exemplificar, dada a sequência de bases nitrogenadas AATAAG, buscamos um códon de parada (TAA, TGA, TAG) através da leitura da sequência a partir da primeira base (da esquerda para a direita), dividindo a sequência em duas partes: AAT e AAG. Para essa primeira leitura, nenhum códon de parada foi encontrado. Então é realizada uma segunda leitura, iniciando a divisão na segunda base, separando assim a sequência em três partes: A, ATA, AG (as sequências que não formam um trio de bases nitrogenadas são ignoradas). Portanto, para essa leitura, também não foi encontrado nenhum códon de parada. Enfim, seguimos para a terceira leitura, dividindo a sequência também em três partes: AA, TAA, G, e então percebemos um códon de parada na sequência (TAA). O processo é repetido no complemento reverso da sequência de bases nitrogenadas.

Como resultado, essa ferramenta produz uma lista com os genes que alinharam com sequências de nucleotídeos em que foi verificada a existência de códons de parada nos seis *frames* de leitura; além de um arquivo com a distribuição desses genes pelos genomas sob estudo. A lista de genes resultante indica que as sequências de nucleotídeos com as quais foram alinhados não podem corresponder a genes. Ou seja, o gene de referência pode ser considerado exclusivo e não houve erro de anotação.

Essa etapa do projeto foi realizada com a colaboração do ex-aluno do curso de Sistemas de Informação da Universidade de São Paulo, Guilherme Duarte Mattos, em seu Trabalho de Conclusão de Curso. A colaboração se deu no desenvolvimento da ferramenta que implementa a segunda tarefa da análise de anotação dos genes exclusivos (fluxo (b) da figura 9).

4.4 Identificação de famílias de genes no próprio genoma

Para verificar a hipótese de que os genes exclusivos podem ter derivado de outras famílias de genes do próprio organismo, foi implementada uma etapa de análise das sequências de nucleotídeos do próprio genoma do gene exclusivo. O objetivo desta etapa da análise é identificar quais genes exclusivos alinharam com outras sequências semelhantes, pertencentes a outras famílias de genes, dentro do próprio genoma.

A ferramenta implementada nessa etapa analisa os alinhamentos dos genes realmente exclusivos identificados na etapa de análise dos erros de anotação (descrita na seção 4.3) com a sequência completa de nucleotídeos do seu respectivo genoma original.

O fluxo da ferramenta desenvolvida está apresentado na figura 10.

Figura 10 – Análise de genes exclusivos alinhados com outras famílias de genes dentro do próprio genoma.

Conforme apresentado na figura 10, a primeira tarefa dessa etapa é identificar os dois melhores alinhamentos do genoma para cada gene exclusivo. Sabe-se que, os alinhamentos de cada genoma são ordenados pelo parâmetro *score*, do maior para o menor. Além disso, sabe-se que o primeiro alinhamento (o melhor) é o alinhamento do gene com sua própria sequência de nucleotídeos. Isto posto, foram selecionados os dois melhores alinhamentos: o da própria sequência e o melhor alinhamento com outra família de genes, ou seja, a família de genes que mais se assemelha com o gene em questão.

A partir dessa seleção de famílias de genes alinhadas aos genes exclusivos, foram separados os genes que possuíam apenas um alinhamento após a seleção, ou seja, os genes que alinharam apenas com eles mesmos. Esses genes não possuem nenhuma semelhança relevante com outras famílias de genes dentro do próprio genoma.

A seguir, foram identificados os valores de identidade e cobertura dos alinhamentos. A identidade do alinhamento é obtida pelo terceiro atributo do arquivo de alinhamento: *percent identities*. Já a cobertura do alinhamento é obtida a partir do cálculo da subtração da posição de fim da sequência (*query sequence end*) pela posição inicial (*query sequence start*), dividida pelo tamanho do alinhamento ($([sequenceEnd - (sequenceStart + 1)] / alignmentSize)$). O tamanho do alinhamento, por sua vez é obtido pelo tamanho do alinhamento do gene com ele mesmo (*alignment length*).

Por fim, os genes que apresentaram alinhamentos relevantes (de acordo com a identidade e cobertura do alinhamento) com outras famílias de genes foram agrupados em faixas de acordo com esses valores de identidade e cobertura. Cada faixa apresenta 10% do total da porcentagem de identidade ou cobertura, ou seja, o primeiro grupo contém alinhamentos com identidade e cobertura entre 90% e 100%, o segundo grupo está entre 80% e 90% de identidade e entre 70% e 80% de cobertura. Tais grupos foram criados a fim de analisar esses alinhamentos, identificando graus de semelhança dos genes exclusivos e as famílias de genes alinhadas a eles.

4.5 Taxonomia de genes exclusivos

Para testar a hipótese de que as sequências dos genes exclusivos se diferenciaram muito das sequências originais dos ancestrais, foi implementada uma etapa de análise das sequências de nucleotídeos dos genes exclusivos em relação a outras espécies com genoma

sequeenciado. O objetivo desta etapa é identificar quais genes exclusivos alinharam com sequências de outros organismos (não necessariamente anotadas como genes), a fim de identificar um potencial ancestral comum.

A ferramenta implementada nessa etapa analisa os alinhamentos das sequências de nucleotídeos dos genes considerados exclusivos identificados na etapa de análise dos erros de anotação (descrita na seção 4.3) com as sequências completas de outras espécies do banco do NCBI. O fluxo da ferramenta desenvolvida está apresentado na figura 11.

Figura 11 – Análise de genes exclusivos alinhados com sequências de nucleotídeos de outras espécies do banco do NCBI.

Fonte: Priscilla Koch Wagner, 2018

Conforme apresentado na figura 11, a ferramenta recebe um arquivo com os alinhamentos das sequências de nucleotídeos dos genes do genoma em estudo, alinhado com sequências de nucleotídeos do banco do NCBI. O primeiro passo é filtrar apenas os

alinhamentos de genes exclusivos, a partir da lista de genes exclusivos identificados na etapa de análise de erros de anotação (seção 4.3).

Após obtidos os alinhamentos de interesse (sequências dos genes exclusivos com sequências do banco do NCBI), alguns parâmetros de filtro utilizados pelo algoritmo que identifica a taxonomia são definidos: coberturaMIN, identidadeMIN e evaluateMAX. CoberturaMIN indica a cobertura mínima que um alinhamento deve possuir para ser considerado pelo algoritmo de taxonomia. IdentidadeMIN indica o valor mínimo de identidade do alinhamento para ser considerado pelo algoritmo de taxonomia. Da mesma forma que foi feito na etapa de análise de famílias de genes (seção 4.4), os parâmetros coberturaMIN e identidadeMIN variaram de 10% em 10%, ou seja, primeiramente foram considerados alinhamentos com no mínimo 90% de cobertura e 90% de identidade, a seguir, foram considerados alinhamentos com 90% de cobertura e 80% de identidade, e assim por diante, até que o valor de cobertura e identidade mínima chegassem em zero.

O parâmetro evaluateMAX indica o valor máximo do parâmetro e-value que o alinhamento deve ter para ser considerado pelo algoritmo. Conforme apresentado na seção 4.1, o e-value indica a probabilidade do alinhamento ter ocorrido por coincidência, quanto menor seu valor, menores são as chances do alinhamento ter ocorrido ao acaso. A variação desse parâmetro utilizou dois valores: 0 e 10^{-10} , sendo 0 o valor mais restritivo e 10^{-10} o valor menos restritivo.

Após a definição desses parâmetros, foi executado o algoritmo que identifica a taxonomia da espécie de interesse a partir dos genes exclusivos identificados. O algoritmo de taxonomia funciona da seguinte maneira: primeiramente são identificadas as sub-árvores de cada espécie, com os níveis taxonômicos mais baixos, a partir dos alinhamentos. A seguir, as árvores vão sendo conectadas a partir da identificação do nível taxonômico comum acima, até ser identificado o nível taxonômico mais alto. A taxonomia das espécies utilizada neste mestrado é a provida pelo NCBI. Cada genoma do banco é associado a uma espécie ou patovar dessa taxonomia.

O algoritmo de taxonomia tem como retorno uma árvore com os níveis taxonômicos do genoma de interesse, bem como os níveis taxonômicos das demais espécies com as quais o genoma de interesse foi alinhado (de acordo com os parâmetros definidos em cada execução). Em cada nível taxonômico, são identificados quantos genes foram identificados nesse nível, por exemplo, dentro do nível *Xanthomonas* (gênero) de uma árvore taxonômica podem existir N genes, que são distribuídos entre as espécies no nível seguinte.

5 Experimentos e Resultados

A partir dos estudos realizados na área de biologia comparativa, análise filogenética de genes exclusivos e técnicas de análise de dados biológicos e genéticos, foi definido um conjunto de experimentos visando a investigar as hipóteses levantadas e apresentadas no capítulo 4, bem como a análise filogenética de genes exclusivos.

Para esses experimentos, foram consideradas as etapas de análise filogenética, sendo aplicadas ao problema específico dos genes exclusivos de espécies próximas.

Este capítulo contém a descrição dos dados utilizados, experimentos realizados e resultados obtidos durante esse projeto de pesquisa para análise de genes exclusivos. Inicialmente, os dados utilizados nos experimentos estão descritos na seção 5.1. A seguir, o experimento realizado no estudo de caso está apresentado na seção 5.2. A seção 5.3 apresenta a descrição dos experimentos. Por fim, os resultados da análise de erros de anotação, da análise dos alinhamentos com outras famílias de genes no próprio genoma e da análise dos genes exclusivos com outras espécies estão apresentados nas seções 5.4, 5.5 e 5.6, respectivamente.

5.1 Descrição dos Dados

Para os experimentos realizados durante o desenvolvimento deste trabalho, optou-se por utilizar genomas sequenciados de espécies de bactérias do gênero *Xanthomonas*. Optou-se por utilizar genomas de bactérias, pois, conforme apresentado no capítulo 2, esses organismos têm sido os mais utilizados em pesquisas genéticas por conta do tamanho do seu genoma ser menor do que o genoma de outros seres vivos.

Dentre as bactérias, escolheu-se as bactérias do gênero *Xanthomonas*, pois o melhor entendimento sobre elas pode ajudar a tratar e prevenir doenças em espécies vegetais atacadas por estas bactérias, conforme já apresentado nos capítulos anteriores. Além disso, as bactérias do gênero *Xanthomonas* já estavam sendo utilizadas pelo grupo de pesquisa em que o projeto se insere.

No total, foram escolhidas 15 sub-espécies ou patovares de bactérias do gênero *Xanthomonas*. Por meio da ferramenta BLAST, obteve-se os alinhamentos entre estas 15 sub-espécies, a fim de identificar os genes potencialmente exclusivos.

A tabela 2 apresenta as 15 espécies escolhidas, o número de pares de bases de seus cromossomos e a quantidade total de genes anotados de cada espécie.

Tabela 2 – Genomas utilizados nos experimentos, com respectivas abreviações, seu tamanho e número de genes.

Genoma	Abreviação	Tamanho cromossomo	# de genes
<i>Xanthomonas albilineans</i> GPE PC73	Xalbilineans	3.768.695	3.208
<i>Xanthomonas axonopodis</i> pv. citri str. 306	Xacitri	5.175.554	4.427
<i>Xanthomonas axonopodis</i> pv. citrumelo F1	Xacitrumelo	4.967.469	4.181
<i>Xanthomonas axonopodis</i> Xac29-1	Xac29-1	5.153.455	4.403
<i>Xanthomonas campestris</i> pv. campestris str. 8004	Xcc8004	5.148.708	4.271
<i>Xanthomonas campestris</i> pv. campestris str. ATCC 33913	XccATCC33913	5.076.188	4.179
<i>Xanthomonas campestris</i> pv. campestris str. B100	XccB100	5.079.002	4.466
<i>Xanthomonas campestris</i> pv. raphani 756C	Xraphani	4.941.214	4.516
<i>Xanthomonas campestris</i> pv. vesicatoria str. 85-10	Xvesicatoria	5.178.466	4.726
<i>Xanthomonas citri</i> subsp. citri Aw12879	Xccitri	5.321.499	4.760
<i>Xanthomonas fuscans</i> subsp. fuscans	Xff	4.981.995	4.083
<i>Xanthomonas oryzae</i> pv. oryzae KACC 10331	XooKACC10331	4.941.439	4.065
<i>Xanthomonas oryzae</i> pv. oryzae MAFF 311018	XooMAFF311018	4.940.217	4.372
<i>Xanthomonas oryzae</i> pv. oryzae PXO99A	XooPXO99A	5.240.075	4.988
<i>Xanthomonas oryzae</i> pv. oryzicola BLS256	Xooryzicola	4.831.739	4.474

Fonte: Priscilla Koch Wagner, 2018

O objetivo deste trabalho é analisar filogeneticamente os genes exclusivos de quaisquer espécies de genomas próximos. Para isso, durante os experimentos foram utilizados genes potencialmente exclusivos de cada espécie de bactéria *Xanthomonas* escolhida.

Conforme apresentado na seção 4.2 do capítulo anterior, a seleção desses genes potencialmente exclusivos foi realizada a partir de uma ferramenta desenvolvida pelo grupo de pesquisa em que este projeto se insere, após o alinhamento entre as sub-espécies de *Xanthomonas*.

A partir disso, temos a quantidade de genes potencialmente exclusivos, ou seja, que não foram alinhados com nenhum outro gene das 15 espécies próximas selecionadas (considerando os limiares utilizados para os parâmetros dos alinhamentos). A tabela 3 apresenta a relação da quantidade total de genes potencialmente exclusivos pela quantidade total de genes de cada genoma. Genes potencialmente exclusivos são aqueles que estavam anotados nos arquivos oriundos do NCBI e que não possuem nenhum alinhamento (dentro dos limiares restritivos utilizados) com outro gene anotado.

Nota-se que, para a maioria das espécies, mais de 90% dos genes foram considerados potencialmente exclusivos, pois não são semelhantes a nenhum outro gene de outros genomas próximos considerados.

No total, as 15 espécies somam 65.119 genes. Os genes que apresentaram algum alinhamento com algum gene de outro genoma podem ser visualizados na figura 12, obtida

Tabela 3 – Relação entre a quantidade total de genes exclusivos e a quantidade total de genes de cada genoma

Abreviação	# total de genes	# de genes potencialmente exclusivos	% total exclusivos
Xalbilineans	3.208	3.162	98,57%
Xacitri	4.427	640	14,46%
Xacitrumelo	4.181	3.296	78,83%
Xac29-1	4.403	4.344	98,66%
Xcc8004	4.271	1.213	28,40%
XccATCC33913	4.179	4.153	99,38%
XccB100	4.466	4.390	98,30%
Xraphani	4.516	4.513	99,93%
Xvesicatoria	4.726	4.697	99,39%
Xccitri	4.760	4.743	99,64%
Xff	4.083	4.080	99,93%
XooKACC10331	4.065	2.296	56,48%
XooMAFF311018	4.372	4.149	94,90%
XooPXO99A	4.988	4.960	99,44%
Xooryzicola	4.474	4.474	100,00%

Fonte: Priscilla Koch Wagner, 2018

do trabalho de Pereira (2017). Essa figura apresenta uma rede de genes, relacionando os genes dos 15 genomas em estudo. O foco do trabalho de Pereira (2017) está em estudar as famílias de genes homólogos apresentados na rede da figura 12, enquanto o foco deste trabalho está em estudar os genes restantes, aprimorando a estratégia de identificação dos genes exclusivos.

Figura 12 – Rede de genes das 15 espécies de *Xanthomonas* em análise neste estudo, apresentando as relações entre as diferentes famílias de genes homólogos dos 15 genomas.

Fonte: Pereira (2017), p.74.

A partir dos genes homólogos dos 15 genomas das espécies de *Xanthomonas* é possível gerar uma matriz de distância e, a partir dessa matriz, um cladograma (ou árvore filogenética). A figura 13, também obtida do trabalho de Pereira (2017), apresenta o cladograma que representa a filogenia de referência obtida utilizando os mesmos 15 genomas deste estudo. Para a geração dessa visualização, foram considerados todos os genes de todos os genomas das espécies.

Figura 13 – Filogenia de Referência

Fonte: Pereira (2017), p.74.

A figura 13 apresenta as relações de semelhanças entre os genomas das espécies analisadas neste capítulo.

5.2 Estudo de Caso

Esta seção descreve o experimento e resultados do estudo de caso realizado no início do desenvolvimento deste projeto, descrito no capítulo anterior. O presente estudo de caso utilizou genes exclusivos da bactéria da espécie *Xanthomonas campestris pv. campestris*. O objetivo deste experimento foi validar a proposta deste projeto de pesquisa e demonstrar sua aplicabilidade no problema de geração de filogenia de genes exclusivos.

Conforme descrito no capítulo anterior, foram identificados os genes exclusivos da bactéria *Xanthomonas campestris pv. campestris* utilizando uma ferramenta desenvolvida pelo grupo de pesquisa (PEREIRA, 2017). Esses genes foram alinhados com os genes de organismos do banco de genomas do NCBI, utilizando a ferramenta BLAST (ALTSCHUL et al., 1990). No total, foram identificados 136.256 alinhamentos.

A partir desses alinhamentos, foram realizadas as primeiras análises. A primeira análise manual mostrou que os genes de referência alinharam apenas com genes da própria espécie *Xanthomonas campestris pv. campestris*, ou seja, os genes exclusivos da bactéria *Xanthomonas campestris pv. campestris* não estão presentes em nenhum outro genoma do

banco do NCBI. A partir dessa informação, surgiram as hipóteses descritas no capítulo 4, e que foram exploradas pelos experimentos descritos nas seções seguintes deste capítulo.

Os alinhamentos identificados pela ferramenta BLAST foram utilizados em uma análise taxonômica. Foi utilizada a ferramenta descrita na seção 4.5 do capítulo anterior. Como não existiam alinhamentos com outras espécies, esse experimento apresentou apenas a taxonomia da espécie dos genes de referência *Xanthomonas campestris pv. campestris*. A ferramenta de taxonomia apresenta alguns filtros que selecionam os alinhamentos relevantes para a análise taxonômica dos genes de referência. Os filtros utilizados para esse estudo de caso foram:

- Porcentagem de identidade (mínima): 80%;
- Comprimento alinhamento (mínimo): 60;
- *Mismatches* (máximo): 20;
- *Gap openings* (máximo): 5;
- *Bit score* (mínimo): 100;
- *E-value* (máximo): 10^{-10} .

Com os filtros indicados acima, a ferramenta de taxonomia filtrou 741 alinhamentos relevantes dos 136.256 obtidos através da ferramenta BLAST. A taxonomia da espécie *Xanthomonas campestris pv. campestris* está apresentada na figura 14.

Figura 14 – Taxonomia da espécie *Xanthomonas campestris pv. campestris*, considerando 741 alinhamentos com o genoma da própria espécie.

```
cellular organisms [741]
  Bacteria [741]
    Proteobacteria [741]
      Gammaproteobacteria [741]
        Xanthomonadales [741]
          Xanthomonadaceae [741]
            Xanthomonas [741]
              Xanthomonas campestris [741]
                Xanthomonas campestris pv. campestris [741]
```

Este estudo de caso valida a proposta do projeto, sendo aplicável ao problema de geração de filogenia de genes exclusivos. A partir dos resultados apresentados, referentes à bactéria *Xanthomonas campestris pv. campestris*, é possível concluir que os genes exclusivos utilizados são um primeiro passo para o estudo de comportamentos específicos dessa espécie de *Xanthomonas*.

A identificação de alinhamentos de genes exclusivos apenas com o próprio genoma, em si já é um bom resultado, pois a partir disso foi possível identificar hipóteses de estudo que foram exploradas durante este trabalho e trouxeram importantes considerações sobre o problema de análise filogenética de genes exclusivos. Além disso, a partir deste estudo de caso, identificou-se a oportunidade de realizar experimentos variando os parâmetros da ferramenta de taxonomia, a fim de avaliar alinhamentos considerados menos relevantes em análises filogenéticas genéricas, mas que podem ser importantes na análise de genes exclusivos.

5.3 Descrição dos Experimentos

Todos os experimentos foram realizados usando as ferramentas descritas no capítulo 4 e consideraram os genomas das espécies apresentadas na seção 5.1 deste capítulo. Cada grupo de experimentos visou a analisar um sub-problema da abordagem proposta, e foram divididos da seguinte maneira:

- Experimentos para identificar erros de anotação: utilizou as ferramentas apresentadas na seção 4.3. Os dados de alinhamentos das 15 espécies de *Xanthomonas* foram obtidos utilizando a ferramenta BLAST, alinhando todos os genes (ou sequências de nucleotídeos) de cada organismo com todos os genes (ou sequências de nucleotídeos) dos demais e com filtro de *e-value* máximo igual a 10^{-10} (parâmetro padrão utilizado na literatura para homologias, como por exemplo em Kihara (2011)). A partir desses arquivos de alinhamento obtidos, foram geradas as matrizes de presença de genes e de nucleotídeo, filtrando os alinhamentos relevantes com os seguintes parâmetros (obtidos via uma ferramenta desenvolvida pelo grupo de pesquisa para maximizar o coeficiente de clusterização da rede de genes):
 - identidade mínima de 96%;
 - cobertura mínima de 96%;
 - tamanho mínimo do alinhamento de 60 bases;
 - número máximo de 20 erros;
 - número máximo de 5 aberturas de *gaps*;
 - *e-value* máximo de 10^{-10} ;
 - *score* mínimo de 100.

Para a segunda etapa deste experimento, foi identificada qual sequência que melhor se alinhou com um gene considerado inicialmente exclusivo, exceto o próprio gene. Esses alinhamentos foram filtrados com os parâmetros:

- identidade mínima de 96%;
- cobertura mínima de 96%;

Além desses alinhamentos, também foram utilizados, na segunda etapa deste experimento, os arquivos no formato FASTA, descritos na seção 4.1 do capítulo 4. Esses arquivos possuem as sequências de nucleotídeos dos genomas de referência das 15 espécies utilizadas nos alinhamentos.

- Experimentos com alinhamentos com o próprio genoma: utilizou a ferramenta apresentada na seção 4.4. Os dados de entrada utilizados foram os 15 arquivos (um para cada genoma de interesse) de alinhamentos, dos genes considerados realmente exclusivos identificados na análise de anotação descrita na seção 5.4 com os genes do próprio genoma, obtidos a partir da ferramenta BLAST. O parâmetro de *e-value* máximo do BLAST foi suavizado para este experimento: 10^{-5} .
- Experimentos com alinhamentos com outras espécies (NCBI): utilizou a ferramenta apresentada na seção 4.5. Como dados de entrada, foram utilizados 15 arquivos de alinhamentos, um para cada genoma de interesse, gerados a partir da ferramenta BLAST. Cada arquivo de alinhamento contém alinhamentos entre os genes exclusivos de uma das 15 espécies de *Xanthomonas* e genes do banco de genomas NCBI. O parâmetro *e-value* utilizado neste experimento teve valor igual a 10^{-5} .

5.4 Análise de Erros de Anotação

Este primeiro experimento tem o objetivo de validar a hipótese de que existem possíveis inconsistências nas anotações dos genes dos organismos estudados. Como visto na seção 5.1 deste capítulo, a maior parte dos 15 genomas utilizados nos experimentos apresenta mais de 90% dos genes sendo considerados potencialmente exclusivos. Isso corrobora a hipótese da existência dos erros de anotação nos genomas estudados.

Para este experimento, foram utilizadas as ferramentas apresentada na seção 4.3 do capítulo 4. Na primeira tarefa, o objetivo foi analisar quais genes anotados em cada genoma apresentavam sequências de nucleotídeos não anotados nos outros genomas próximos. Para

isso, foram utilizadas matrizes de presença de genes e de nucleotídeos dos 15 genomas de *Xanthomonas*. Cada matriz de presença de genes indica quantas vezes um gene aparece em cada genoma do estudo, enquanto a matriz de presença de nucleotídeos indica quantas vezes a sequência de nucleotídeos correspondente a um gene (ou uma CDS, região codificante do gene) de um genoma aparece em outro genoma (não necessariamente anotado como gene).

Conforme descrito na seção 4.3, em relação a um gene, podem existir erros (ou inconsistências) de anotação dentro do próprio genoma (intragenoma) e erros de anotação em relação a outros genomas (intergenoma). Esse experimento considera ambos tipos de erros para os genes considerados potencialmente exclusivos inicialmente.

As quantidades de genes potencialmente exclusivos por genoma estão apresentadas na tabela 3 da seção 5.1, sendo um total de 55.110 genes potencialmente exclusivos para os 15 genomas em análise.

Ao analisar os erros intragenoma dos genes potencialmente exclusivos das 15 espécies de *Xanthomonas* em estudo, foram identificados 5.142 genes que apresentaram essa inconsistência, 9,33% dos 55.110 considerados para as 15 espécies. A distribuição de genes com erros intragenoma em cada espécie considerada está apresentada na figura 15.

Figura 15 – Distribuição de genes com erros intragenoma entre as 15 espécies de *Xanthomonas* consideradas.

Conforme apresentado na figura 15, dos 5.142 genes com erro (ou inconsistência) de anotação intragenoma identificados nas 15 espécies, 40,63% foram identificados na espécie *Xalbilineans*. O segundo genoma com mais erros intragenoma em relação ao total de erros das 15 espécies é o *Xooryzicola*, com 11,73% do total de genes com esse tipo de inconsistência nas 15 espécies. Estes erros ou inconsistências costumam ocorrer pelo fato dos genomas terem seus potenciais genes identificados e anotados em épocas diferentes e utilizando metodologias diferentes. Uma das formas que vários grupos de pesquisa têm utilizado para contornar este tipo de problema é executar um novo processo de identificação e anotação de genes (utilizando uma única metodologia) de todos os genomas em estudo.

Analisando os erros ou inconsistências intragenoma em relação ao total de genes considerados potencialmente exclusivos inicialmente em cada espécie, vemos que, dos 3.162 genes potencialmente exclusivos da *Xalbilineans*, 66,07% apresentaram alguma inconsistência. O segundo genoma com maior quantidade de genes com este erro em relação aos potencialmente exclusivos inicialmente é o *Xff*, com 13,48% (4.474 potencialmente exclusivos inicialmente).

Já é sabido pela comunidade científica que estuda esses genomas que a *Xalbilineans* possui grandes divergências em relação aos demais genomas próximos de *Xanthomonas*. Tais divergências podem ser causadas pela própria anotação do genoma, visto a quantidade de genes anotados com inconsistências dentro dele próprio.

Ao contabilizar a quantidade de potenciais erros de anotação intragenoma em cada genoma, chegou-se ao total de 5.318, apresentados na figura 16.

Figura 16 – Espécies com erros intragenoma a mais que a quantidade de genes com este erro.

No total, existem 176 inconsistências a mais que a quantidade de genes com esse tipo de erro, ou seja, algumas sequências possuem mais de uma cópia não anotada dentro do próprio genoma.

Conforme apresentado na figura 16, 6 dos 15 genomas apresentaram mais de um erro intragenoma por gene. A espécie que apresentou mais desses erros por gene foi a *Xff*, com 69 erros a mais que a quantidade de genes com esse erro, seguida por *Xccitri* com 49 a mais.

Com relação aos erros de anotação intergenoma, ou seja, os que são identificados por sequências de genes não anotadas, semelhantes a um gene de uma espécie, encontradas em genomas de outras espécies, a figura 17 apresenta a distribuição de genes com esse tipo de inconsistência.

Figura 17 – Distribuição de genes envolvidos em potenciais erros intergenoma entre as 15 espécies de *Xanthomonas* consideradas.

Fonte: Priscilla Koch Wagner, 2018

No total, foram encontrados 28.996 genes envolvidos nesse tipo de inconsistência, 52,61% dos 55.110 considerados para as 15 espécies. Conforme apresentado na figura 17, 10 dos 15 genomas estudados apresentaram mais de dois mil genes envolvidos em erros ou inconsistências de anotação em outros genomas. Destaca-se que os problemas potenciais de anotação não ocorrem necessariamente com todos os 28.996 genes identificados, o que estes números informam é que há genes homólogos a eles não identificados no próprio genoma ou em algum dos outros genomas em estudo que apresentam essa inconsistência

de identificação/anotação. Por exemplo, dado um gene que possui uma cópia em cada um dos 15 genomas estudados.: Se uma cópia desse gene já estava inicialmente identificada e anotada corretamente em 14 genomas, mas não em um dos genomas, então estes 15 genes estão sendo computados como envolvidos em uma inconsistência de anotação dentro do grupo de genomas estudados.

Novamente, a *Xalbilineans* se destaca das demais espécies, apresentando apenas dois genes com erros de anotação em outros genomas, o que representa 0,01% do total de genes encontrados com essa inconsistência. A espécie *Xacitri* também teve comportamento diferente das demais, com apenas 100 genes com erros de anotação em outros genomas, correspondendo a 0,34% do total. Em contrapartida, a espécie com mais erros desse tipo é a *Xccitri*, com 2.799 genes com inconsistência de anotação, representando 9,65%.

Ao analisar a relação entre os genes envolvidos nesse tipo inconsistência em cada espécie e o total de genes considerados potencialmente exclusivos inicialmente, a maioria das espécies estudadas apresentou mais de 50% de erros, como apresentado na figura 18. Isto quer dizer que, ao se analisar grupos de genes homólogos entre estes genomas (considerando apenas as sequências fornecidas como genes), uma quantidade muito grande de grupos estaria incompleta (ou incorreta) devido a estas inconsistências de anotação identificadas.

Figura 18 – Relação entre os genes que apresentaram erro intergenoma e o total de genes considerados potencialmente exclusivos inicialmente.

Fonte: Priscilla Koch Wagner, 2018

A espécie que possui a maior quantidade de genes com erros ou inconsistências de anotação em outros genomas em relação ao total de genes considerados potencialmente exclusivos é a *XccATCC33913*, com 66,48%. A *Xacitri*, que apresentou um índice muito

baixo de erros em relação a quantidade total de erros intergenoma encontrados, quando analisada em relação a quantidade de genes potencialmente exclusivos, esse índice muda significativamente para 15,65%, visto que esse genoma possui um total de 640 genes considerados potencialmente exclusivos.

A figura 19 apresenta a distribuição dos potenciais erros de anotação intergenoma pela quantidade de genomas em que o erro ocorreu (isto é, quantos genomas possuem genes homólogos àqueles envolvidos em potenciais erros de anotação). Por exemplo, se um gene apresentou um erro em relação a sete genomas distintos, ele está computado na coluna 7 do gráfico.

Figura 19 – Distribuição de erros de anotação por quantidade de genomas em que o erro ocorreu.

Fonte: Priscilla Koch Wagner, 2018

Com base na figura 19, nota-se que 36,73% dos genes que apresentaram erros de anotação, possuíam sequências semelhantes em 3 genomas distintos. Uma quantidade significativa de 2.806 genes apresentaram erros em 13 genomas distintos, correspondendo a 9,68% do total de genes com erros intergenoma.

A figura 20 apresenta a frequência com que cada genoma teve seus genes envolvidos em erros ou inconsistências de anotação em relação aos demais, considerando os erros ou inconsistências presentes em 3 genomas (10.650 no total, como apresentado na figura 19).

Conforme é possível identificar na figura 20, o genoma *Xcc8004* possui 5.538 sequências semelhantes a genes presentes em algum dos outros 14 genomas. Esses 5.538

Figura 20 – Frequência que cada genoma esteve envolvido em erros ou inconsistências de anotação em relação aos demais, considerando erros presentes em 3 genomas.

Fonte: Priscilla Koch Wagner, 2018

genes possuem homólogos em outros dois genomas (pois no total cada gene indicado estava presente em três genomas). Esse valor corresponde a 52% do total de genes que apresentaram erros ou inconsistências de anotação com três genomas. Novamente, a espécie *Xalbilineans* é a que apresentou menos genes envolvidos em erros ou inconsistências em relação a outros genomas, apenas 10 sequências semelhantes, o que corresponde a 0,09% do total de 10.650 genes que apresentaram erros de anotação com 3 genomas.

A figura 21 apresenta um diagrama, chamado diagrama de *Chord*, onde é possível visualizar as inter-relações entre dois genomas, sendo um o genoma de origem, que possui o gene anotado e o outro o genoma em que o respectivo gene não está anotado.

No diagrama de *Chord* da figura 21, os nós representam os genomas e são dispostos ao longo de um círculo, com arcos que fazem relações entre eles. Os arcos representam a quantidade de erros que existem de um genoma para o outro, sendo a quantidade de erros proporcional ao tamanho de cada arco.

Observando o diagrama é possível notar que o padrão se mantém, sendo a *Xalbilineans*, a *Xacitri* e a *Xcc8004* as espécies com menos erros (5, 479 e 4.556, respectivamente), enquanto as demais apresentam uma quantidade elevada de erros. A maior quantidade de

Figura 21 – Diagrama de *Chord* apresentando as inter-relações entre os genomas que possuem os genes anotados e os que apresentaram o erro.

Fonte: Priscilla Koch Wagner, 2018

erros de um genoma para outro está entre *XccATCC33913* e *Xcc8004*, com 2.743 genes com erros.

A espécie *XccATCC33913* apresenta 3 genomas com os quais possui uma maior quantidade de relacionamentos. A figura 22 apresenta um diagrama de *Chord* considerando apenas a espécie *XccATCC33913* como genoma de origem, a fim de facilitar a visualização dos erros relacionados a essa espécie.

A espécie *XccATCC33913* apresentou mais erros ou inconsistências em relação aos genomas das espécies *Xcc8004*, com 2.743 erros; *XccB100*, com 2.522 erros e *Xraphani*, com 2.340 erros (25,53%, 23,47% e 21,78% do total de erros da espécie, respectivamente). Os demais genomas apresentaram menos de 400 erros cada um (número ainda significativo), em relação a espécie *XccATCC33913*.

A tabela 4 apresenta a comparação de genes exclusivos antes e após a análise, a relação entre a quantidade de genes de cada genoma e a quantidade de realmente exclusivos encontrados na análise e a taxa de diminuição da quantidade de genes realmente exclusivos em relação aos potenciais exclusivos identificados inicialmente.

Figura 22 – Diagrama de *Chord* apresentando os erros intergenoma entre a espécie *XccATCC33913* e os que apresentaram os erros.

Fonte: Priscilla Koch Wagner, 2018

Tabela 4 – Comparação de genes exclusivos antes e após a análise, relação da quantidade de genes do genoma com os exclusivos encontrados e diminuição da quantidade de genes exclusivos.

Abreviação	# total de genes	# de exclusivos antes da análise	# de exclusivos após análise	% de genes exclusivos	# de diminuição de exclusivos
<i>Xalbilineans</i>	3.208	3.162	3.160	98,50%	0,06%
<i>Xacitri</i>	4.427	640	540	12,20%	2,26%
<i>Xacitrumelo</i>	4.181	3.296	1.446	34,59%	44,25%
<i>Xac29-1</i>	4.403	4.344	1.588	36,07%	62,59%
<i>Xcc8004</i>	4.271	1.213	703	16,46%	11,94%
<i>XccATCC33913</i>	4.179	4.153	1.392	33,31%	66,07%
<i>XccB100</i>	4.466	4.390	1.780	39,86%	58,44%
<i>Xraphani</i>	4.516	4.513	2.129	47,14%	52,79%
<i>Xvesicatoria</i>	4.726	4.697	2.132	45,11%	54,27%
<i>Xccitri</i>	4.760	4.743	1.944	40,84%	58,80%
<i>Xff</i>	4.083	4.080	1.951	47,78%	52,14%
<i>XooKACC10331</i>	4.065	2.296	1.030	25,34%	31,14%
<i>XooMAFF311018</i>	4.372	4.149	1.548	35,41%	59,49%
<i>XooPXO99A</i>	4.988	4.960	2.319	46,49%	52,95%
<i>Xooryzicola</i>	4.474	4.474	2.452	54,81%	45,19%

Fonte: Priscilla Koch Wagner, 2018

A quantidade final total de genes considerados exclusivos após essa análise foi de 26.114. Isso corresponde a 47,39% da quantidade de genes considerados potencialmente exclusivos inicialmente e 40,10% da quantidade total de genes em todos os genomas.

Com base nos dados da tabela 4, a diminuição da quantidade de genes considerados exclusivos inicialmente pela quantidade de genes exclusivos após a análise foi maior que 50% para a maioria dos genomas estudados.

O genoma que mais teve redução na quantidade de exclusivos foi a *XccATCC33913*, com 66,07% de diminuição. A *XccATCC33913* possuía 4.153 genes potencialmente exclusivos e após a análise ficou com 1.392 realmente exclusivos (33,31% do total de genes do genoma).

Como notado no decorrer das análises, a *Xalbilineans* foi a espécie com menor redução da quantidade de exclusivos. Apenas dois de seus genes apresentaram erros de anotação e foram desconsiderados na contagem de exclusivos, apresentando uma diminuição de apenas 0,06%.

Essa informação por si só não garante que o genoma da *Xalbilineans* possua poucos erros de anotação. Por se tratar da sub-espécie mais diferente do conjunto selecionado, o seu genoma requer maiores análises em relação a genes exclusivos dentro do gênero de *Xanthomonas*.

Na segunda tarefa da análise de erros de anotação, o objetivo foi analisar as sequências de nucleotídeos que apresentaram erros de anotação com algum gene de outra espécie. Essa análise foi feita a fim de definir se essas sequências podem efetivamente corresponder a genes, corroborando com o resultado da primeira etapa. Ou seja, se a sequência de nucleotídeos for considerada um gene (não tiver códon de parada no meio da sequência), o gene do genoma de origem que alinhou com essa sequência não é exclusivo.

Para isso, foi utilizado um arquivo de alinhamentos de todos os genes anotados alinhados com as sequências de nucleotídeos dos 15 genomas das espécies de *Xanthomonas* em estudo. Além do arquivo de alinhamento, foram usados arquivos FASTA para obter as sequências de nucleotídeos dos alinhamentos de interesse selecionados.

Da análise desses alinhamentos com os arquivos FASTA das sequências de nucleotídeos, foram obtidos 243 alinhamentos que apresentaram sequências de nucleotídeos com a presença de ao menos um códon de parada dentro dos seis quadros de leitura (ou seja, estas sequências de nucleotídeos não podem corresponder a genes). A figura 23 apresenta esses 243 genes alinhados, agrupados por genoma.

É possível notar que, novamente, a *Xalbilineans* não apresentou nenhum alinhamento onde o gene de referência alinhou com alguma sequência com códon de parada. Ou seja, se existirem alinhamentos com sequências de nucleotídeos que não podem codificar genes

Figura 23 – Genes cuja sequência de nucleotídeos com a qual alinhou possui códon de parada.

Fonte: Priscilla Koch Wagner, 2018

para os genes exclusivos da *Xalbilineans*, eles possivelmente não passaram pelos filtros da ferramenta.

Apesar de ser um número relativamente baixo, a *XooMAFF311018* foi a espécie que apresentou maior quantidade de genes exclusivos alinhados a sequências com códon de parada, 61 genes. Esse valor representa 25,10% do total de 243 genes com alinhamentos identificados. Essas sequências de nucleotídeos podem corresponder, por exemplo, a erros no sequenciamento e/ou montagem no genoma ou à degradação do gene.

A fim de desconsiderar os potenciais erros de anotação indicados pela primeira etapa deste processo para os 243 genes que foram confirmados como exclusivos, foi feita a comparação dos genes com erro (28.996 genes) e os genes confirmados como exclusivos após a segunda etapa. Esses erros poderiam ser desconsiderados pois, a sequência com a qual se alinhou não pode codificar um gene.

Porém, nenhum gene considerado exclusivo e que apresentou erro de anotação na primeira etapa do processo estava presente na lista dos 243 genes.

5.5 Análise de alinhamentos com outras famílias de genes no próprio genoma

Este experimento tem o objetivo de validar a hipótese de que os genes exclusivos das *Xanthomonas* derivaram de outras famílias de genes do próprio organismo.

Como visto na seção 5.4 deste capítulo, ao analisar erros de anotação genética dentro do próprio genoma, comparando genes anotados com as sequências de nucleotídeos do genoma completo, foi possível identificar 5.142 erros intragenoma, isso corresponde a 9,33% dos 55.110 genes que haviam sido considerados exclusivos inicialmente nas 15 espécies de *Xanthomonas*.

Essa análise aumenta a chance da hipótese em questão ser correta para parte dos genes considerados exclusivos, pois podem ter se originado de outras famílias de genes, ou sequências de nucleotídeos do próprio genoma em estudo.

Para este experimento, foi utilizada a ferramenta apresentada na seção 4.4 do capítulo 4. O objetivo foi identificar quais e quantos genes exclusivos possuíam semelhanças relevantes com alguma outra família de gene do próprio genoma, podendo ser uma possível origem desse gene exclusivo. Consequentemente será possível saber quais e quantos genes não se originaram do próprio genoma por meio de uma variação de outro gene.

A quantidade de genes exclusivos após a remoção dos erros de anotação foi 26.114, para os 15 genomas em estudo (a quantidade de genes exclusivos por genoma pode ser vista na tabela 4, da seção anterior).

A figura 24 apresenta a quantidade, por espécie, de genes exclusivos que não possuíam nenhum alinhamento com outra família de gene dentro do próprio genoma.

Conforme apresentado na figura 24, 16.929 genes exclusivos dos 15 genomas não alinharam com nenhuma outra família de genes dentro do próprio genoma. Isso equivale a 64,83% do total de genes exclusivos identificados na análise de erros de anotação. A espécie que teve mais genes exclusivos que não se alinharam com nenhuma outra família de gene dentro do próprio genoma foi a *Xalbilineans*, com 1.823 genes.

A figura 25 relaciona os genes que não alinharam com outras famílias do próprio genoma e todos os genes exclusivos para cada espécie de *Xanthomonas*.

Nessa perspectiva, a *Xalbilineans* possui 57,69% de genes exclusivos que não alinharam com outras famílias de genes no próprio genoma. A espécie que teve mais genes não alinhados com outras famílias, em relação ao total de genes exclusivos foi a

Figura 24 – Quantidade, por espécie, de genes exclusivos que não possuíam nenhum alinhamento com outra família de gene dentro do próprio genoma.

Fonte: Priscilla Koch Wagner, 2018

Figura 25 – Relação entre os genes que não alinharam com outras famílias do próprio genoma e todos os genes exclusivos para cada espécie de *Xanthomonas*.

Fonte: Priscilla Koch Wagner, 2018

Xraphani, com 75,01%. Em contrapartida, a espécie em que essa relação foi a menor foi a *XooKACC10331*, com 49,03%.

Os 16.929 genes que não tiveram alinhamentos com outras famílias de gene possivelmente e potencialmente não são originários de outros genes da própria espécie, nem de sequências de nucleotídeos não anotadas em alguma das 15 espécies próximas, portanto, continuam sendo exclusivos.

A figura 26 apresenta o foco desta análise: a quantidade de genes que tiveram algum alinhamento com outras famílias de genes do próprio genoma.

Figura 26 – Quantidade, por espécie, de genes exclusivos que possuem alinhamento com outra família de gene dentro do próprio genoma.

Fonte: Priscilla Koch Wagner, 2018

Conforme apresentado na figura 26, 8.523 genes exclusivos alinharam com outras famílias de gene, considerando os 15 genomas. Esse valor corresponde a 32,64% do total de genes potencialmente exclusivos identificados na análise da sessão anterior. A espécie que teve mais genes que alinharam com outras famílias de genes foi novamente a *Xalbilineans*, com 1.240 genes. Em contrapartida, a espécie que teve menos genes alinhados com outras famílias foi a *Xacitri*, com 162.

A figura 27 relaciona os genes que possuem algum alinhamento com outras famílias do próprio genoma e todos os genes exclusivos para cada espécie de *Xanthomonas*.

Nessa visão relativa, a *Xalbilineans* apresenta 39,24% dos genes exclusivos alinhados com outras famílias de genes do próprio genoma. A espécie que possui mais genes alinhados, em relação ao total de exclusivos foi a *XooKACC10331*, com 50,49%. Em contrapartida, a espécie cujos genes menos alinharam com outras famílias no próprio genoma foi a *Xraphani*, com 24,14%. Esses valores já eram esperados, visto que a relação de genes não alinhados foi inversa em ambas as espécies.

Esses 8.523 genes exclusivos para suas respectivas espécies de *Xanthomonas* provavelmente se originaram de outras famílias de genes do próprio genoma (com as quais alinharam nessas análises), sofrendo algum tipo de mutação no processo.

Figura 27 – Relação entre os genes que possuem alinhamento com outras famílias do próprio genoma e todos os genes exclusivos para cada espécie de *Xanthomonas*.

Fonte: Priscilla Koch Wagner, 2018

As quantidades de alinhamentos de cada genoma, pelos parâmetros de faixa de identidade e faixa de cobertura, estão divididas entre as figuras 28 e 29. Na figura 28 estão apresentados os valores para as espécies: (a) *Xacitri*, (b) *Xacitrumelo*, (c) *Xalbilineans*, (d) *Xac29-1*, (e) *Xcc8004*, (f) *XccATC33913*, (g) *XccB100* e (h) *Xccitri*. Já na figura 29, estão apresentados os valores das espécies: (i) *Xraphani*, (j) *Xvesicatoria*, (k) *Xff*, (l) *XooKACC10331*, (m) *XooMAFF311018*, (n) *XooPXO99A* e (o) *Xooryzicola*.

Como é possível notar nas figuras 28 e 29, com exceção da *Xacitrumelo* (b), *Xalbilineans* (c) e *Xraphani* (i), todas as outras 12 espécies possuem a maior parte dos alinhamentos nas faixas de 100 à 90 de identidade e cobertura. Ou seja, os valores de identidade e cobertura são os mais altos para definir os alinhamentos dessas 12 espécies. No caso da *Xacitrumelo* (b), *Xalbilineans* (c) e *Xraphani* (i), a concentração de alinhamentos se encontra em valores altos de cobertura, mas a faixa de identidade cai bastante.

Outras espécies, apesar de apresentarem a maior quantidade de genomas nas maiores faixas de identidade e cobertura, apresentam uma grande concentração de alinhamentos em faixas mais baixas de identidade. É o caso das espécies *Xac29-1* (d), *XccATC33913* (f), *XccB100* (g), *Xvesicatoria* (j) e *Xff* (k).

As análises das tabelas apresentadas nas figuras 28 e 29 indicam que a grande maioria dos alinhamentos de genes exclusivos com outras famílias de genes do próprio genoma apresentam comprimento igual ou semelhante ao tamanho do gene de referência,

Figura 28 – Quantidade de alinhamentos dos genomas, pelos parâmetros de faixa de identidade e faixa de cobertura. Sendo: (a) *Xacitri*, (b) *Xacitrumelo*, (c) *Xalbilineans*, (d) *Xac29-1*, (e) *Xcc8004*, (f) *XccATC33913*, (g) *XccB100* e (h) *Xccitri*.

Fonte: Priscilla Koch Wagner, 2018

pois a faixa de cobertura da maioria dos alinhamentos é alta. Em contrapartida, na maioria das 15 espécies analisadas, a identidade desses alinhamentos variou bastante.

Figura 29 – Quantidade de alinhamentos dos genomas, pelos parâmetros de faixa de identidade e faixa de cobertura. Sendo: (i) *Xraphani*, (j) *Xvesicatoria*, (k) *Xff*, (l) *XooKACC10331*, (m) *XooMAFF311018*, (n) *XooPXO99A* e (o) *Xooryzicola*.

Fonte: Priscilla Koch Wagner, 2018

5.6 Análise de alinhamentos dos genes exclusivos em outras espécies

O objetivo deste experimento foi tentar identificar possíveis origens de genes exclusivos das 15 espécies de *Xanthomonas* em outras espécies, cruzando esses genes exclusivos com as sequências de nucleotídeos do banco de genomas do NCBI.

Como o estudo de caso apresentado na seção 5.2 mostrou, o alinhamento de genes exclusivos com genes da base do NCBI não trouxe indícios de possíveis origens desses

genes em outras espécies. Uma das hipóteses é que: as sequências se modificaram muito em relação às sequências ancestrais. Outra hipótese é que esses genes exclusivos apresentam sequências semelhantes que não configuram genes (ou não foram anotadas como genes) em outras espécies.

Este experimento utilizou os alinhamentos dos genes exclusivos dos 15 genomas de *Xanthomonas* com o banco de sequências de nucleotídeos do NCBI para gerar árvores taxonômicas de cada uma das espécies, incluindo genomas de espécies com as quais algum(ns) gene(s) exclusivo(s) dos genomas de interesse possuem semelhança.

Os parâmetros de identidade e cobertura para este experimento iniciaram em 100 e variaram de 10 em 10, conforme explicado na seção 4.5 do capítulo 4. Esses parâmetros foram variados a fim de identificar se tal variação gerava árvores taxonômicas diferentes (mais abrangentes, no caso de baixos valores dos parâmetros). O *e-value* também variou: 0 ou 10^{-10} .

Para cada variação de parâmetros foi gerada uma árvore taxonômica a partir dos alinhamentos filtrados. No total, foram geradas 3.000 árvores taxonômicas, sendo 200 para cada espécie (100 para cada valor de *e-value*, com dez faixas de identidade e dez faixas de cobertura).

A fim de analisar a árvore taxonômica mais abrangente e menos abrangente de cada espécie, todas as árvores taxonômicas foram comparadas. Com isso, foram identificadas as árvores mais e menos abrangentes, e os parâmetros que geraram essas árvores. A abrangência foi definida pela quantidade de espécies que fazem parte da árvore taxonômica. As tabelas 5 e 6 apresentam os dados das árvores menos abrangentes e mais abrangentes selecionadas, respectivamente.

As taxonomias selecionadas, apresentadas nas tabelas 5 e 6, foram escolhidas visando obter a menor e a maior árvore taxonômica para cada uma das espécies. Ou seja, dado o mínimo de alinhamentos necessários, com os parâmetros mais restritivos, obtivemos as árvores com menos níveis taxonômicos e, conseqüentemente, menos espécies relacionadas. Ao mesmo tempo, dadas as mesmas condições, obtivemos as árvores com maiores níveis taxonômicos. Para boa parte dos genomas, a mudança de cobertura não apresentou muita alteração em relação ao tamanho da árvore taxonômica gerada, mas apresentou diferença em relação à quantidade de alinhamentos utilizados para geração da árvore. Foram selecionadas as taxonomias que usaram menos alinhamentos para obter um mesmo resultado.

Tabela 5 – Parâmetros e resultados das árvores menos abrangente para as 15 espécies de *Xanthomonas*.

Espécie	Taxonomia menos abrangente				
	<i>e-value</i>	Identidade	Cobertura	# alinhamentos	# espécies envolvidas
<i>Xalbilineans</i>	10^{-10}	100-90	90-80	6	2
<i>Xacitri</i>	10^{-250}	100-90	100-90	106	1
<i>Xacitrumelo</i>	10^{-250}	100-90	100-90	10712	3
<i>Xac29-1</i>	10^{-250}	100-90	100-90	21972	4
<i>Xcc8004</i>	10^{-250}	100-90	100-90	39233	4
<i>XccATCC33913</i>	10^{-250}	100-90	100-90	62680	4
<i>XccB100</i>	10^{-250}	100-90	100-90	100927	4
<i>Xraphani</i>	10^{-250}	100-90	100-90	163659	4
<i>Xvesicatoria</i>	10^{-250}	100-90	100-90	229963	5
<i>Xccitri</i>	10^{-250}	100-90	100-90	336317	5
<i>Xff</i>	10^{-250}	100-90	100-90	448033	5
<i>XooKACC10331</i>	10^{-250}	100-90	100-90	561353	6
<i>XooMAFF311018</i>	10^{-250}	100-90	100-90	676665	6
<i>XooPXO99A</i>	10^{-250}	100-90	100-90	793178	6
<i>Xooryzicola</i>	10^{-10}	90	90	452	1

Fonte: Priscilla Koch Wagner, 2018

Tabela 6 – Parâmetros e resultados das árvores mais abrangente para as 15 espécies de *Xanthomonas*.

Espécie	Taxonomia mais abrangente				
	<i>e-value</i>	Identidade	Cobertura	# alinhamentos	# espécies envolvidas
<i>Xalbilineans</i>	10^{-10}	80-70	90-80	218	3
<i>Xacitri</i>	10^{-10}	100-90	100-90	2103	5
<i>Xacitrumelo</i>	10^{-10}	100-90	100-90	50403	5
<i>Xac29-1</i>	10^{-10}	100-90	100-90	119016	6
<i>Xcc8004</i>	10^{-10}	100-90	100-90	238709	6
<i>XccATCC33913</i>	10^{-10}	100-90	100-90	388707	8
<i>XccB100</i>	10^{-10}	100-90	100-90	668845	8
<i>Xraphani</i>	10^{-10}	100-90	100-90	1111435	9
<i>Xvesicatoria</i>	10^{-10}	100-90	100-90	1584768	11
<i>Xccitri</i>	10^{-10}	100-90	100-90	2259746	11
<i>Xff</i>	10^{-10}	100-90	100-90	2972348	13
<i>XooKACC10331</i>	10^{-10}	100-90	100-90	3705340	15
<i>XooMAFF311018</i>	10^{-10}	100-90	100-90	4477470	16
<i>XooPXO99A</i>	10^{-10}	100-90	100-90	5282568	17
<i>Xooryzicola</i>	10^{-10}	80-70	90-80	41886	3

Fonte: Priscilla Koch Wagner, 2018

Conforme apresentado nas tabelas, todas as taxonomias com maior abrangência apresentaram ao menos 3 espécies, ou seja, pelo menos duas espécies a mais do que a do próprio genoma. A espécie com a taxonomia mais abrangente de todas foi a *XooPXO99A*, com 17 espécies relacionadas. Para essa taxonomia dessa espécie, foram utilizados 5.282.568 alinhamentos.

As únicas espécies cujas taxonomias menos abrangentes não consideraram outras espécies (ou seja, não é possível supor origens de genes), foram a *Xacitri* e *Xooryzicola*. A *Xalbilineans*, mesmo com uma quantidade pequena de genes sendo considerados (apenas 6), apresentou duas espécies relacionadas na taxonomia (uma além da própria espécie).

Em relação aos parâmetros, nota-se que, com exceção de *Xalbilineans* e *Xooryzicola* (apenas para a taxonomia mais abrangente), todos os valores de identidade e cobertura foram máximos para todas as taxonomias relevantes selecionadas. Além disso, todas as taxonomias mais abrangentes utilizaram *e-value* de 10^{-10} , enquanto as menos abrangentes, com exceção de *Xalbilineans* e *Xooryzicola*, utilizaram o *e-value* mais restritivo com valor 0.

A fim de comparar visualmente duas taxonomias, as taxonomias selecionadas da espécie *Xvesicatoria* (espécie escolhida aleatoriamente) estão apresentadas nas figuras 30 e 31.

Figura 30 – Taxonomia menos abrangente da espécie *Xvesicatoria*.

```
cellular organisms [229963]
  Bacteria [227555]
    Proteobacteria [227555]
      Gammaproteobacteria [227519]
        Xanthomonadales [227519]
          Xanthomonadaceae [226013]
            Xanthomonas [226013]
              Xanthomonas citri group [11917]
                Xanthomonas citri [11917]
                Xanthomonas axonopodis [5511]
                Xanthomonas axonopodis Xac29-1 [5511]
                Xanthomonas campestris [91455]
                Xanthomonas campestris pv. campestris [91455]
                Xanthomonas oryzae [601]
                Xanthomonas oryzae pv. oryzae [601]
                Xanthomonas euvesicatoria [12]
                Xanthomonas campestris pv. vesicatoria str. 85-10 [12]
```

Fonte: Priscilla Koch Wagner, 2018

Os dados apresentados nas tabelas 5 e 6, informam os parâmetros mínimos para se obter taxonomias relevantes, para as espécies de *Xanthomonas* considerando alinhamentos de genes exclusivos com sequências de nucleotídeos do NCBI. As taxonomias relevantes relacionam outras espécies de seres vivos, a fim de estudar possíveis origens dos genes considerados exclusivos nesse grupo de espécies próximas. Por exemplo, observa-se pela figura 30 que, utilizando parâmetros mais restritivos, todos os genes considerados exclusivos da *Xanthomonas campestris* pv. *vesicatoria* str. 85-10 possuem alinhamentos apenas com outras quatro espécies de *Xanthomonas*. Por outro lado, ao serem utilizados parâmetros menos restritivos (figura 31) é possível encontrar genes de outros gêneros como *Stenotrophomonas* e *Pseudoxanthomonas* (nestes casos, dois gêneros da família das Xanthomonadaceae).

Figura 31 – Taxonomia mais abrangente da espécie *Xvesicatoria*.

```

cellular organisms [1584768]
  Bacteria [1545839]
    Proteobacteria [1545839]
      Gammaproteobacteria [1447243]
        Xanthomonadales [1447243]
          Xanthomonadaceae [1252616]
            Xanthomonas [1251333]
              Xanthomonas campestris [409483]
                Xanthomonas campestris pv. campestris [409483]
              Xanthomonas oryzae [5888]
                Xanthomonas oryzae pv. oryzae [5888]
              Xanthomonas citri group [16824]
                Xanthomonas citri [16824]
              Xanthomonas axonopodis [23547]
                Xanthomonas axonopodis Xac29-1 [23547]
              Xanthomonas euvesicatoria [21]
                Xanthomonas campestris pv. vesicatoria str. 85-10 [21]
              Xanthomonas albilineans [702]
                Xanthomonas albilineans GPE PC73 [702]
              Stenotrophomonas [780]
                Stenotrophomonas maltophilia group [780]
                  Stenotrophomonas maltophilia [780]
              Pseudoxanthomonas [1]
              Lysobacter [502]
            Betaproteobacteria [301]
              Burkholderiales [301]
                Comamonadaceae [301]
                  Acidovorax [301]
  environmental samples [101]

```

Fonte: Priscilla Koch Wagner, 2018

As ferramentas desenvolvidas possibilitam que os genes exclusivos e respectivas árvores filogenéticas sejam analisados caso a caso, de acordo com o interesse ou a necessidade do pesquisador de biologia comparativa.

Vale ressaltar, como resultado deste experimento, a possibilidade de criar filogenias mais abrangentes (que consideram mais espécies para compor a taxonomia) usando parâmetros de identidade e cobertura mais restritivos (maiores valores, ou seja, que trazem os alinhamentos mais relevantes) para a maioria das espécies de *Xanthomonas*. Foi possível notar durante esse experimento que usando parâmetros mais restritivos foi possível obter, na maioria dos casos, a mesma taxonomia mais abrangente que usando os parâmetros mais abrangentes, porém, com um número muito menor de alinhamentos considerados, o que torna o processamento do experimento mais rápido, por conta da quantidade de alinhamentos considerada para ser analisada ser menor.

Outro resultado interessante é que para a maioria das espécies de *Xanthomonas* foi possível encontrar taxonomias que relacionam várias espécies distantes evolutivamente,

dando evidências de que a hipótese de que os genes exclusivos atuais dessas espécies podem ter se diferenciado dos genes ancestrais comuns, mas as sequências de nucleotídeos ainda existentes nas espécies mais distantes continuam sendo semelhantes aos genes exclusivos dessas espécies de *Xanthomonas*.

6 Conclusão

Este trabalho teve como principal objetivo desenvolver uma estratégia, através do desenvolvimento de ferramentas computacionais, para lidar com o problema de identificação e análise filogenética de genes exclusivos de genomas próximos evolutivamente, auxiliando na identificação de possíveis origens de tais genes. Durante análises iniciais, identificou-se a necessidade de, primeiramente, analisar as anotações genômicas dos genes exclusivos, dada a hipótese de que haviam inconsistências nessas anotações. A partir dessa necessidade de análise prévia de anotações genômicas, foi possível analisar com maior precisão a origem dos genes considerados realmente exclusivos, removendo aqueles que apresentaram erros de anotação.

Com base nas análises de potenciais erros de anotação, foi possível identificar um comportamento diferenciado de alguns dos genomas estudados. Por exemplo, conforme esperado, a *Xalbilineans* se mostrou a espécie que mais se difere das demais espécies de *Xanthomonas* estudadas, apresentando poucos erros ou inconsistências de anotação em relação às outras espécies, porém, com inúmeras sequências de nucleotídeos não anotadas dentro do próprio genoma que se assemelham a genes considerados exclusivos (em uma comparação de genes de referência com genes de outras espécies). Essas sequências possivelmente são a origem dos genes exclusivos dessa espécie (ou se originaram deles), sendo possível supor que boa parte dos genes exclusivos atuais da *Xalbilineans* surgiram de outras sequências ou genes do próprio genoma, evidenciando assim sua grande divergência em relação às demais espécies. Outra espécie que se destacou nessa análise foi a *Xcc8004*, que mais esteve envolvida em potenciais erros de anotação de genes exclusivos de outras espécies. Isso pode sugerir que as anotações genômicas dessa espécie foram feitas utilizando um método diferente do utilizado nas demais espécies ou em uma época diferente e seria interessante que fossem reanalisadas. Por fim, pode-se concluir a partir dessas análises que existem muitas inconsistências nas anotações genômicas de bactérias do gênero *Xanthomonas* (e possivelmente de outras espécies de seres vivos), que podem causar interpretações errôneas se utilizadas para construções de análises filogenéticas dessas espécies. Desta forma, uma prática que vem sendo utilizada por vários grupos de pesquisa que trabalham com genômica comparativa é recomendada: a reanotação (identificação dos

genes e suas prováveis funções) de todos os genomas que serão analisados utilizando um mesmo método.

A partir do experimento que visa analisar alinhamentos de genes exclusivos de espécies de *Xanthomonas* com outras famílias de genes dentro do próprio genoma, novamente, a espécie *Xalbilineans* se destacou, apresentando maior quantidade de genes com alinhamentos relevantes dentro do próprio genoma. Tal resultado corrobora a análise descrita no parágrafo anterior, que supõe como possível origem dos genes exclusivos da *Xalbilineans* as sequências ou genes do próprio genoma. Ainda a partir deste experimento, foi possível notar que a maioria dos alinhamentos de boa parte das espécies de *Xanthomonas* apresentou alinhamentos com valores máximos (dentre os testados) dos parâmetros de identidade e cobertura. Isso significa que as famílias de genes com as quais esses genes exclusivos foram alinhados possuem grande semelhança genética com os genes em questão, sendo possível supor que podem ser originários uns dos outros. Por fim, pode-se concluir com as análises deste experimento que a hipótese de que os genes exclusivos podem ter se originado de famílias de genes do próprio genoma se mostrou bastante provável para boa parte dos genes exclusivos das espécies consideradas.

Com base no experimento que analisou alinhamentos de genes exclusivos com outras espécies mais distantes evolutivamente, pode-se concluir que foram encontradas relações relevantes para serem estudadas caso a caso por pesquisadores da área de biologia comparativa. Um resultado interessante deste experimento foi a possibilidade de criar filogenias mais abrangentes (que consideram mais espécies para compor a taxonomia) usando parâmetros de identidade e cobertura mais restritivos (maiores valores, ou seja, que trazem os alinhamentos mais relevantes) para a maioria das espécies de *Xanthomonas*. Durante a seleção das taxonomias mais e menos abrangentes, notou-se que usando parâmetros mais restritivos foi possível obter, na maioria das vezes, a mesma taxonomia mais abrangente que usando os parâmetros mais abrangentes, porém, com um número muito menor de alinhamentos considerados, na maioria dos casos. Isso torna o processamento do experimento mais rápido, visto que a quantidade de alinhamentos considerada para ser analisada é menor. Por fim, para a maioria das espécies de *Xanthomonas* foi possível encontrar taxonomias que relacionam várias espécies distantes evolutivamente, dando evidências de que a hipótese de que os genes exclusivos atuais dessas espécies podem ter se diferenciado dos genes ancestrais comuns, mas as sequências de nucleotídeos ainda existentes nas espécies mais distantes continuam sendo semelhantes aos genes exclusivos dessas espécies.

de bactérias *Xanthomonas* ou mesmo que ocorreu um processo de transferência lateral (hipótese que não foi estudada a fundo no presente trabalho).

Além disso, é importante salientar que, apesar dos experimentos realizados utilizarem um conjunto de genomas de bactérias do gênero *Xanthomonas*, as ferramentas desenvolvidas durante este projeto podem ser utilizadas para realizar análises filogenéticas de genes exclusivos de genomas de outras espécies. Os resultados obtidos a partir dos experimentos com genomas de *Xanthomonas* indicam que as estratégias desenvolvidas e implementadas podem ser úteis para identificar erros de anotação ou possíveis origens de genes exclusivos de outras espécies próximas evolutivamente, que não foram consideradas neste trabalho. Adicionalmente, a maioria das ferramentas desenvolvidas é parametrizável, permitindo a um especialista no domínio calibrar as ferramentas de acordo com a proximidade filogenética das espécies em estudo ou de outros objetivos específicos.

As próximas seções apresentam de forma resumida as principais contribuições (seção 6.1) obtidas neste trabalho e os possíveis trabalhos futuros (seção 6.2) a serem desenvolvidos a partir dos resultados obtidos neste trabalho ou para complementar os estudos de genes exclusivos.

6.1 Principais contribuições

As principais contribuições obtidas por meio deste trabalho foram:

- revisão sistemática sobre ferramentas utilizadas para etapas de análise filogenética, publicada como resumo em (WAGNER; PEREIRA; DIGIAMPIETRI, 2015);
- desenvolvimento de algoritmos para identificação e análise de potenciais erros de anotação de genes;
- desenvolvimento de algoritmos para análise de possíveis origens de genes exclusivos dentro do próprio genoma ou em genomas de outras espécies mais distantes evolutivamente;
- adaptação em algoritmo de geração de taxonomia a partir de um conjunto de genes, para torná-lo aplicável ao problema de análise filogenética de genes exclusivos;
- experimento para identificação de genes realmente exclusivos e erros de anotação em bactérias do gênero *Xanthomonas*, com base em alinhamentos de genes e de

sequências de nucleotídeos, publicado em (WAGNER; MATTOS; DIGIAMPIETRI, 2017);

- experimento para identificação de possíveis origens de genes exclusivos de bactérias do gênero *Xanthomonas* em outras famílias de genes do próprio genoma;
- experimento para construção de taxonomias de espécies de bactérias do gênero *Xanthomonas*, baseada em seus genes exclusivos;
- contribuições para a área de biologia comparativa sobre genes exclusivos das espécies de *Xanthomonas* estudadas.

Além da revisão sistemática realizada durante este projeto, também houve colaboração na revisão sistemática conduzida por outra aluna do grupo de pesquisa, a respeito de comparação de genomas completos de bactérias, cujos resultados obtidos foram publicados como resumo em (PEREIRA; WAGNER; DIGIAMPIETRI, 2015)

6.2 Trabalhos Futuros

Como trabalhos futuros, pode-se citar a realização de estudos que envolvam um número e variedade maior de espécies e sub-espécies do gênero *Xanthomonas*, a fim de analisar os erros de anotação em relação a mais espécies, visto que podem ser encontradas outras sequências não anotadas em outros genomas não considerados neste trabalho. Além disso, será possível identificar genes exclusivos das demais espécies desse gênero de bactérias.

Outro estudo que pode ser realizado é considerando outros gêneros de seres vivos, a fim de analisar a aplicabilidade da abordagem proposta em outros genomas mais complexos que os das bactérias, a fim de trazer benefícios em outras frentes.

Para pesquisadores da área de biologia e genética, podem ser aproveitados os resultados sobre os erros de anotação. Um trabalho que pode ser realizado nessa área é a análise em laboratório das sequências de nucleotídeos que não estavam anotadas como genes, mas são semelhantes a um gene considerado exclusivo inicialmente em uma das espécies, a fim de identificar se tal sequência corresponde a uma sequência codificante, ou seja, se a sequência é capaz de codificar proteínas.

Pretende-se, por fim, publicar também os resultados obtidos sobre possíveis origens de genes exclusivos em um periódico da área.

Referências¹

- ALMEIDA, N. F. et al. Comparison of genomes using high-performance parallel computing. In: *Proceedings. 15th Symposium on Computer Architecture and High Performance Computing*. [S.l.: s.n.], 2003. p. 142–148. Citado na página 42.
- ALTSCHUL, S. F. et al. Basic local alignment search tool. *Journal of Molecular Biology*, v. 215, n. 3, p. 403–410, Oct 1990. Citado 5 vezes nas páginas 28, 44, 47, 55 e 66.
- AMORIM, D. d. S. *Fundamentos de Sistemática Filogenética*. Ribeirão Preto: Holos, 2002. 156 p. Citado na página 14.
- BORGES, D. L. et al. Detecting and grading severity of bacterial spot caused by *xanthomonas* spp. in tomato (*solanum lycopersicon*) fields using visible spectrum images. *Comput. Electron. Agric.*, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 125, n. C, p. 149–159, jul. 2016. Citado na página 42.
- BRAMMER, G.; WILLIAMS, T. Using decision trees to study the convergence of phylogenetic analyses. In: *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2010 IEEE Symposium on*. [S.l.: s.n.], 2010. p. 1–8. Citado 3 vezes nas páginas 33, 35 e 39.
- BRENNER, D. J.; STALEY, J. T.; KRIEG, N. R. Classification of procaryotic organisms and the concept of bacterial speciation. In: *Bergey's Manual® of Systematic Bacteriology: Volume Two*. Boston: Springer US, 2005. p. 27–32. Citado na página 29.
- CHEN, A. C. H.; HSIAO, H. C. W.; TSAI, J. J. P. Phylogenetic analysis using nuclear-encoded mitochondrial proteins. In: *Proceedings of the 2009 Ninth IEEE International Conference on Bioinformatics and Bioengineering*. Washington, DC, USA: IEEE Computer Society, 2009. (BIBE '09), p. 374–377. Citado 5 vezes nas páginas 33, 34, 36, 37 e 39.
- DIXIT, R. et al. Brief communication: Molecular and phylogenetic analysis of a novel family of fibrinogen-related proteins from mosquito *aedes albopictus* cell line. *Comput. Biol. Chem.*, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 32, n. 5, p. 382–386, out. 2008. Citado 4 vezes nas páginas 33, 35, 36 e 39.
- FANG, X.-Y.; LUO, Z.-G.; WANG, Z.-H. Predicting rna secondary structure using profile stochastic context-free grammars and phylogenetic analysis. *J. Comput. Sci. Technol.*, Institute of Computing Technology, Beijing, China, v. 23, n. 4, p. 582–589, jul. 2008. Citado 4 vezes nas páginas 33, 34, 35 e 39.
- FITCH, W. *Distinguishing Homologous from Analogous Proteins*. [S.l.]: Syst Zool, 1970. 99-113 p. Citado na página 22.
- FREEMAN, S.; HERRON, J. C. *Análise Evolutiva*. [S.l.]: Artmed Editora, 2009. Citado 4 vezes nas páginas 22, 25, 27 e 30.
- GERTZ, E. M. et al. Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. *BMC Biology*, v. 4, n. 1, p. 1, 2006. Citado na página 29.

¹ De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

- HALGASWATHTHA, T. et al. Neural network based phylogenetic analysis. In: *Biomedical Engineering (ICoBE), 2012 International Conference on*. [S.l.: s.n.], 2012. p. 155–160. Citado 3 vezes nas páginas 33, 35 e 39.
- HILLIS, D. M.; MORITZ, C.; MABLE, B. *Molecular Systematics*. [S.l.]: Sinauer Associates, 1996. Citado na página 23.
- HOEF-EMDEN, K. Molecular phylogenetic analyses and real-life data. *Computing in Science Engineering*, v. 7, n. 3, p. 86–91, May 2005. Citado 2 vezes nas páginas 31 e 32.
- HU, Y. et al. A simple method for phylogenetic analysis of dna sequences. In: *Computer Science and Network Technology (ICCSNT), 2013 3rd International Conference on*. [S.l.: s.n.], 2013. p. 139–143. Citado 3 vezes nas páginas 33, 38 e 39.
- JACQUES, M.-A. et al. Using ecology, physiology, and genomics to understand host specificity in *Xanthomonas*. *Annual Review of Phytopathology*, v. 54, p. 163–187, 2016. Citado na página 24.
- JALAN, N. et al. Comparative genomic and transcriptome analyses of pathotypes of *xanthomonas citri* subsp. *citri* provide insights into mechanisms of bacterial virulence and host range. In: *BMC Genomics*. [S.l.: s.n.], 2013. Citado 2 vezes nas páginas 24 e 41.
- KIHARA, D. *Protein Function Prediction for Omics Era*. Springer Netherlands, 2011. (SpringerLink : Bücher). ISBN 9789400708815. Disponível em: <https://books.google.com.br/books?id=UPUZy11Y4c8C>. Citado na página 68.
- KITCHENHAM, B. *Procedures for Performing Systematic Reviews*. Department of Computer Science, Keele University, UK, 2004. Citado 2 vezes nas páginas 17 e 31.
- LIMA-JUNIOR, J. D. et al. Characterization of mycobacteria and mycobacteriophages isolated from compost at the são paulo zoo park foundation in brazil and creation of the new mycobacteriophage cluster u. *BMC Microbiology*, v. 16, n. 1, p. 111, Jun 2016. ISSN 1471-2180. Disponível em: <https://doi.org/10.1186/s12866-016-0734-3>. Citado na página 24.
- MATIOLI, S. R. *Biologia Molecular e Evolução*. Ribeirão Preto: Editora Holos, 2001. 202 p. Citado 4 vezes nas páginas 14, 20, 21 e 22.
- MAZET, F. et al. Phylogenetic relationships of the fox (forkhead) gene family in the bilateria. *Gene*, v. 316, n. 0, p. 79 – 89, 2003. ISSN 0378-1119. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0378111903007418>. Citado na página 100.
- MISHRA, A. et al. Phylogenetic analysis and molecular modeling of wheat pathogenesis-related proteins: Serpin and peroxidase. In: *Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference on*. [S.l.: s.n.], 2015. p. 1139–1143. Citado 4 vezes nas páginas 33, 36, 37 e 39.
- MIYAMOTO, M.; CRACRAFT, J. *Phylogenetic Analysis of DNA Sequences*. [S.l.]: Oxford University Press, 1991. Citado 4 vezes nas páginas 14, 24, 26 e 30.
- MOREIRA, L. M. (Org.). *Ciências genômicas : fundamentos e aplicações*. 1. ed. Ribeirão Preto: Sociedade Brasileira de Genética - SBG, 2015. Citado 4 vezes nas páginas 25, 26, 29 e 45.

MOREIRA, L. M. et al. Novel insights into the genomic basis of citrus canker based on the genome sequences of two strains of *Xanthomonas fuscans* subsp. *aurantifolii*. *BMC Genomics*, v. 11, p. 238–262, 2010. Citado 4 vezes nas páginas 15, 23, 40 e 41.

NILSSON, R. H.; LARSSON, K.-H.; URSING, B. M. galaxie—cgi scripts for sequence identification through automated phylogenetic analysis. *Bioinformatics*, Oxford University Press, Oxford, UK, v. 20, n. 9, p. 1447–1452, jun. 2004. Citado 6 vezes nas páginas 33, 34, 35, 36, 37 e 39.

NOGUEIRA, C. L. et al. Characterization of mycobacterium chelonae-like strains by comparative genomics. *Frontiers in Microbiology*, v. 8, p. 789, 2017. Citado 2 vezes nas páginas 40 e 44.

OCHMAN, H.; LAWRENCE, J. G.; GROISMAN, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature*, v. 405, n. 6784, p. 299–304, 2000. Citado na página 23.

PANCHAL, H.; ANJARIA, K.; PATEL, S. Phylogenetic analysis of some leguminous trees using clustalw2 bioinformatics tool. In: *Proceedings of the 2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*. Washington, DC, USA: IEEE Computer Society, 2012. (BIBMW '12), p. 917–921. Citado 4 vezes nas páginas 33, 35, 36 e 39.

PEREIRA, V. M. Y. *Reconstrução filogenética de procariotos com base em famílias de genes homólogos*. Dissertação (Mestrado) — Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, 2017. Citado 4 vezes nas páginas 44, 64, 65 e 66.

PEREIRA, V. M. Y.; WAGNER, P. K.; DIGIAMPIETRI, L. A. Bacterial whole genome comparison: a systematic literature review. In: *X-Meeting 2015 - 11th International Conference of the AB3C + Brazilian Symposium of Bioinformatics*. [S.l.: s.n.], 2015. Citado na página 94.

REQUENO, J. I. et al. Temporal logics for phylogenetic analysis via model checking. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, IEEE Computer Society Press, Los Alamitos, CA, USA, v. 10, n. 4, p. 1058–1070, jul. 2013. Citado 2 vezes nas páginas 33 e 38.

RUDOLPH, K. Infection of the plant by *Xanthomonas*. In: *Xanthomonas*. London: Chapman & Hall, 1993. p. 193–264. Citado na página 24.

RYAN, R. P. et al. Pathogenomics of *xanthomonas*: understanding bacterium-plant interactions. In: *Nature Reviews Microbiology*. [S.l.: s.n.], 2011. p. 344–355. Citado 2 vezes nas páginas 24 e 41.

SCHAECHTER, M. (Org.). *Encyclopedia of Microbiology*. 3. ed. Oxford: Academic Press, 2009. Citado na página 29.

SETUBAL, J. C.; MEIDANIS, J. *Introduction to Computational Molecular Biology*. Boston, EUA: PWS Publishing Company, 1997. 296 p. Citado 11 vezes nas páginas 14, 20, 21, 24, 25, 26, 27, 28, 29, 30 e 100.

TAN, S.; WU, S. A web-based phylogenetic analysis for a query gene. In: *Bioinformatics and Biomedical Engineering, (iCBBE) 2011 5th International Conference on*. [S.l.: s.n.], 2011. p. 1–4. Citado 6 vezes nas páginas 33, 35, 36, 37, 39 e 40.

TARAWNEH, M. et al. Investigation of relationships among excavate taxa using a new quartet-based phylogenetic analysis method. In: *Proceedings of the 21st International Conference on Advanced Information Networking and Applications Workshops - Volume 01*. Washington, DC, USA: IEEE Computer Society, 2007. (AINAW '07), p. 662–667. Citado 2 vezes nas páginas 33 e 38.

TIAN, C. et al. Phylogenetic analysis of members of the genus vibrios based on gyrB genes and 16s rRNA genes. In: *Bioinformatics and Biomedical Engineering, 2008. ICBBE 2008. The 2nd International Conference on*. [S.l.: s.n.], 2008. p. 136–139. Citado 5 vezes nas páginas 33, 35, 36, 37 e 39.

VERLI, H. (Org.). *Bioinformática: da biologia à flexibilidade molecular*. 1. ed. São Paulo: Sociedade Brasileira de Bioquímica e Biologia Molecular - SBBq, 2014. Citado 3 vezes nas páginas 27, 28 e 29.

WAGNER, P. K.; MATTOS, G. D.; DIGIAMPIETRI, L. Análise de erros de anotação de genes exclusivos. In: *XI Brazilian e-Science Workshop (BreSci 2017)*. [S.l.]: Anais do XXXVII Congresso da Sociedade Brasileira de Computação (CSBC 2017), 2017. Citado na página 94.

WAGNER, P. K.; PEREIRA, V. M. Y.; DIGIAMPIETRI, L. A. A systematic review of phylogenetic analysis of a specific gene. In: *X-Meeting 2015 - 11th International Conference of the AB3C + Brazilian Symposium of Bioinformatics*. [S.l.: s.n.], 2015. Citado na página 93.

WANG, J. et al. A new distance computing method for DNA sequences in phylogenetic analysis. In: *Fuzzy Systems and Knowledge Discovery (FSKD), 2013 10th International Conference on*. [S.l.: s.n.], 2013. p. 690–694. Citado 3 vezes nas páginas 33, 38 e 39.

XIA, J. et al. Support vector machine method on predicting resistance gene against *Xanthomonas oryzae* pv. *oryzae* in rice. *Expert Syst. Appl.*, Pergamon Press, Inc., Tarrytown, NY, USA, v. 37, n. 8, p. 5946–5950, ago. 2010. Citado na página 42.

ZAHA, A.; FERREIRA, H. B.; PASSAGLIA, L. M. P. *Biologia Molecular Básica*. Porto Alegre: Editora Mercado Aberto, 2003. 424 p. Citado 3 vezes nas páginas 21, 22 e 23.

ZHANG, S.; WANG, T. A novel alignment-free method for phylogenetic analysis of protein sequences. In: *Proceedings of the 10th WSEAS International Conference on Applied Computer Science*. Stevens Point, Wisconsin, USA: World Scientific and Engineering Academy and Society (WSEAS), 2010. (ACS'10), p. 67–71. Citado 4 vezes nas páginas 33, 36, 37 e 39.

ZHANG, Y. et al. Positive selection is the main driving force for evolution of citrus canker-causing *Xanthomonas*. In: *ISME Journal*. [S.l.: s.n.], 2015. Citado 2 vezes nas páginas 24 e 42.

ZHANG, Z. et al. Phylogenetic analysis of members of the order Laminariales based on rbcL genes and 5.8S rRNA genes. In: *Bioinformatics and Biomedical Engineering, 2009. ICBBE 2009. 3rd International Conference on*. [S.l.: s.n.], 2009. p. 1–3. Citado 5 vezes nas páginas 33, 35, 36, 37 e 39.

ZHOU, T. et al. An approach for determining evolutionary distance in network-based phylogenetic analysis. In: *Proceedings of the 4th International Conference on Bioinformatics Research and Applications*. Berlin, Heidelberg: Springer-Verlag, 2008. (ISBRA'08), p. 38–49. Citado 3 vezes nas páginas 33, 37 e 39.

Apêndice A – Protocolo para Revisão Sistemática: Estado da Arte em análise filogenética de um gene específico

Objetivos:

Identificar e analisar os métodos e técnicas existentes para realizar análise filogenética de um dado gene, considerando genomas do banco mundial e que usem ferramentas de inteligência artificial e/ou ferramentas de bioinformática.

Questões de Pesquisa:

1. 1. Quais são os métodos e técnicas existentes para fazer a análise filogenética de um dado gene?

Intervenção: métodos e técnicas de análise filogenética e classificação genética.

Controle: artigos, livros, teses sobre análise filogenética e classificação genética obtidos com o orientador. Tais artigos utilizados como controle foram: (MAZET et al., 2003), (SETUBAL; MEIDANIS, 1997).

População: projetos de análise filogenética que consideram genomas do banco mundial e que usem ferramentas de inteligência artificial e/ou ferramentas de bioinformática.

Resultados: visão geral do estado da arte da área; e conhecimento abrangente dos métodos e técnicas de análise filogenética para construção de filogenias baseada em genes.

Aplicação: pesquisadores da área de bioinformática que desenvolvem ferramentas de análise filogenética ou classificação genética ou que usem tais ferramentas em suas pesquisas.

Seleção de fontes:

As fontes deverão estar disponíveis via web, preferencialmente em bibliotecas digitais científicas da área. Poderão ser selecionados também, trabalhos disponíveis em outros meios, desde que atendam aos requisitos da Revisão Sistemática.

Palavras-chave:

Será utilizada a expressão “*phylogenic analysis*” ou “*phylogenetic analysis*”, combinada aos termos: *tool*, *algorithm*, “*computational method*”, *framework*, *infrastructure* e demais sinônimos (quando a base de dados não for específica da área de computação).

Listagem de fontes:

- Biblioteca Digital do IEEE (<http://ieeexplore.ieee.org/Xplore/>)
- Biblioteca Digital da ACM (<http://portal.acm.org/>)
- Biblioteca Digital da PLOS One (<http://plosone.org>)
- Base de dados Science Direct (<http://sciencedirect.com>)
- Base de dados PubMed (<http://www.ncbi.nlm.nih.gov/pubmed/>)
- Base de dados Scopus (<http://scopus.com>)

Tipo dos artigos:

Serão considerados artigos completos da área de bioinformática publicados em eventos ou revistas com revisão por pares.

Idioma(s) dos artigos:

Inglês, pois é internacionalmente aceita para trabalhos científicos na área de bioinformática.

*Critérios de inclusão e exclusão dos trabalhos:***Critérios de inclusão:**

- (a) Serão incluídos trabalhos publicados e disponíveis integralmente em bases de dados científicas.
- (b) Serão incluídos os trabalhos que abordarem métodos de identificação da filogenia a partir de um ou mais genes;

- (c) Serão incluídos os trabalhos que usem ferramentas de inteligência artificial e/ou ferramentas de bioinformática para análise filogenética.
- (d) Serão incluídos os trabalhos que abordarem conceitos ou revisões sobre análise filogenética.

Critérios de exclusão:

- (a) Serão excluídos trabalhos que apresentem métodos que não sejam para análise filogenética;
- (b) Serão excluídos trabalhos que não usam ferramentas de inteligência artificial ou de bioinformática para análise filogenética.
- (c) Serão excluídos trabalhos que não estejam disponíveis integralmente nas bases de dados científicas ou que não estejam disponíveis em inglês.

Critérios de qualidade dos estudos primários:

O trabalho deverá ter sido publicado em periódico ou anais de eventos da área de bioinformática com revisão por outros pesquisadores (revisão por pares). Para avaliar os artigos serão utilizados os seguintes critérios:

- (a) possui conhecimento novo/útil para a área de análise filogenética; (peso 3)
- (b) apresenta avaliação do método/técnica apresentado e critérios de validação adotados; (peso 3)
- (c) apresenta precisão e clareza nos resultados; (peso 2)
- (d) permite replicação; (peso 2)
- (e) qualis do veículo de publicação. (peso 2)
- (f) quantidade e confiabilidade das referências; (peso 1)

Processo de seleção dos estudos primários:

Serão construídas strings com as palavras-chave e seus sinônimos. As strings serão submetidas às máquinas de busca das fontes citadas. Para os artigos obtidos pelas máquinas de busca, será feita a leitura do resumo e aplicação dos critérios de inclusão e exclusão. Em

seguida, o trabalho será selecionado se confirmada a sua relevância pelo principal revisor. Se houver dúvida da relevância os demais revisores serão consultados.

Estratégia de extração de informação:

Após definidos os trabalhos definitivamente incluídos, estes serão lidos na íntegra. O revisor fará um resumo de cada um deles, destacando os métodos e técnicas utilizados na análise filogenética.

Serão preenchidos “formulários de extração de dados” para cada texto, considerado válido para a revisão sistemática, lido integralmente. Tais formulários precisam conter as informações básicas:

- Autores;
- País da instituição em que os autores estão filiados;
- Ano da publicação;
- Tipo de publicação;
- Base de dados onde o artigo foi encontrado;
- Abstract.

Além disso, outras características relevantes para a revisão sistemática devem ser extraídos a partir da leitura integral deles:

- Técnica utilizada;
- Tarefa de análise filogenética abordada;
- Base de dados;
- Critério de validação;
- Síntese e análise do trabalho.

Sumarização dos resultados

Após a leitura e o resumo dos trabalhos selecionados, será elaborado um artigo com uma análise quantitativa e qualitativa dos trabalhos a fim de definir as vantagens e desvantagens de cada método. Também serão identificadas e descritas possíveis lacunas no

estado da arte de análise filogenética de genes específicos que podem ser exploradas em pesquisas futuras.