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

VIVIAN MAYUMI YAMASSAKI PEREIRA

**Reconstrução filogenética de procariotos com base em famílias de genes
homólogos**

São Paulo

2017

VIVIAN MAYUMI YAMASSAKI PEREIRA

**Reconstrução filogenética de procariotos com base em famílias de genes
homólogos**

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 03 de abril de 2017. 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

2017

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)

Pereira, Vivian Mayumi Yamassaki

Reconstrução filogenética de procariotos com base em famílias de genes homólogos / Vivian Mayumi Yamassaki Pereira ; orientador, Luciano Antonio Digiampietri. – São Paulo, 2017.
112 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. I. Digiampietri, Luciano Antonio, orient. II. Título

CDD 22.ed.– 572.80285

Dissertação de autoria de Vivian Mayumi Yamassaki Pereira, sob o título “**Reconstrução filogenética de procariotos com base em famílias de genes homólogos**”, 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 03 de abril de 2017 pela comissão julgadora constituída pelos doutores:

Prof. Dr. Luciano Antonio Digiampietri

Universidade de São Paulo

Presidente

Prof. Dr. Marcio Katsumi Oikawa

Universidade Federal do ABC

Prof. Dr. André Fujita

Universidade de São Paulo

Profa. Dra. Arianne Machado Lima

Universidade de São Paulo

Dedico este trabalho aos meus queridos pais, Maria e Primo, e irmã, Larissa, pelo apoio e por tudo o mais que têm feito por mim, não só durante o mestrado como em todos os outros momentos de minha vida.

Agradecimentos

Agradeço, primeiramente, aos meus pais e irmã por toda a compreensão e incentivo que me deram e que me permitiram concluir mais esta etapa.

Ao meu orientador, Prof. Dr. Luciano Antonio Digiampietri, por sua solicitude, por todos os conselhos e orientações dados desde a iniciação científica realizada na graduação e por me mostrar o quanto a bioinformática e o estudo de bactérias podem ser interessantes e igualmente desafiadores.

Ao grupo de pesquisa, composto por Caio, Geraldo, Giovani e Priscilla, por me auxiliarem em várias etapas deste projeto.

Às amigas e também mestrandas Bárbara, Jéssica e Priscilla, pela ajuda e pelas angústias e alegrias que compartilhamos durante o mestrado.

Aos professores do Programa de Pós-graduação em Sistemas de Informação (PPgSI) por contribuírem com a minha formação. Agradeço especialmente à Profa. Dra. Sarajane Marques Peres por toda a ajuda e por ter me incentivado a ingressar no mestrado.

Por fim, gostaria de agradecer à CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), pelo apoio financeiro, e à Universidade de São Paulo (USP).

*“The sure and definite determination (of each species of bacteria) requires so much time,
so much acumen of eye and judgment, so much perseverance and patience that there is
hardly anything so difficult.”*
(MÜLLER, 1786)

Resumo

PEREIRA, Vivian Mayumi Yamassaki. **Reconstrução filogenética de procariotos com base em famílias de genes homólogos**. 2017. 112 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2017.

A comparação de genomas é uma importante tarefa na qual a bioinformática pode ser aplicada, uma vez que ela permite a identificação de genes patogênicos, o que, por sua vez, pode auxiliar a combater ou a prevenir o surgimento de doenças. A partir da comparação de genomas, também é possível realizar a análise filogenética, que permite entender as relações evolutivas entre diferentes organismos. Em genomas de bactérias, essa análise geralmente é realizada com base no gene *16S rRNA*. Entretanto, apesar de ser amplamente utilizado, filogenias com base nesse gene podem ter dificuldades para diferenciar organismos muito próximos evolutivamente. Essa importância da comparação de genomas e a necessidade de uma metodologia que permita distinguir organismos evolutivamente próximos na análise filogenética motivaram este trabalho, que teve como objetivo implementar ferramentas computacionais para identificar genes homólogos em genomas e, com base nesses genes, gerar filogenias e analisar se é possível distinguir os organismos evolutivamente próximos nessas filogenias. Para tanto, as ferramentas desenvolvidas para identificação de genes homólogos recebem resultados de alinhamentos e os filtram, de modo que dois genes são considerados homólogos se o alinhamento entre eles satisfizer os limiares definidos. Após a identificação das famílias de genes homólogos, tabelas são geradas com informações a respeito dos genes homólogos em cada genoma e, com base nessas tabelas, é possível gerar matrizes de distância e utilizar métodos de agrupamento hierárquico para a geração da filogenia ou realizar alinhamentos múltiplos com os genes identificados para posterior reconstrução filogenética. Além disso, também é possível representar os genes e famílias de genes homólogos por meio de um grafo, que pode auxiliar na escolha dos limiares para filtrar os alinhamentos. Para demonstrar e analisar a aplicabilidade das ferramentas desenvolvidas e das abordagens adotadas, experimentos foram realizados utilizando genomas de bactérias do gênero *Xanthomonas*, que contém um grande grupo de bactérias que causam doenças em plantas. Os resultados obtidos foram então comparados com filogenias de referência e com resultados de outros experimentos realizados. Essas comparações demonstraram que as famílias de genes homólogos podem ser úteis para distinguir genomas de organismos muito próximos evolutivamente, apesar de que essa abordagem apresentou dificuldades para separar os grupos de genomas mais distantes. Em contrapartida, na filogenia gerada a partir da região *16S rRNA*, foi possível diferenciar esses organismos mais distantes, mas não foi possível distinguir os organismos muito próximos. Por fim, os experimentos realizados fornecem indícios de que as ferramentas desenvolvidas e as abordagens adotadas podem ser úteis para diferenciar genomas muito próximos evolutivamente de outros procariotos além das bactérias estudadas neste trabalho.

Palavras-chaves: Bioinformática. Genômica comparativa. Genes homólogos. Filogenia.

Abstract

PEREIRA, Vivian Mayumi Yamassaki. **Phylogenetic reconstruction of prokaryotes based on homologous gene families**. 2017. 112 p. Dissertation (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2017.

Genome comparison is an important task on which bioinformatics can be used because it allows the identification of pathogen genes which can aid the combat of diseases and to avoid the emerging of new ones. Genome comparison also allows the phylogenetic analysis which provides the understanding of evolutionary relations of different organisms. In bacterial genomes, this analysis is commonly based on *16S rRNA* gene. Unfortunately, it can present some difficulties to distinguish closely related organisms. This importance of genome comparison and the necessity of a methodology to distinguish organisms that are closely related motivated this study, which aimed the development of computational tools to identify homologous genes in genomes, to use these genes to reconstruct phylogenies and to analyze if it is possible to distinguish closely related organisms on these phylogenies. To achieve this purpose, the developed tools to identify homologous genes receive the alignments results and filter it, such that two genes are homologous if their alignment satisfies the thresholds. After the identification of homologous gene families, the tools generates tables with information about the homologous genes presents in each genome and with these tables it is possible to create distance matrix to be used by hierarchical clustering methods to generate phylogenies or it is possible to perform multiple alignments with the identified genes to accomplish a phylogenetic reconstruction. Besides that, it is possible to represent the genes and homologous gene families in a graph, which can aid the choice of the thresholds to filter the alignments. To demonstrate and analyze the applicability of the developed tools and the approaches chosen in this study, experiments were performed using genomes of the bacterial genus *Xanthomonas*, which include a group of phytopathogenic bacteria. The results obtained were compared with reference phylogenies and with results of other experiments. These comparisons showed that homologous gene families can be used to differentiate closely related organisms, despite the fact that it presented difficulties to distinguish the groups of genomes that were evolutionarily far from each other. On the other hand, the phylogeny based on *16S rRNA* region allows to distinguish the groups of genomes that were distant, but it was not possible to differentiate closely related organisms. As a conclusion, the experiments performed give pieces of evidence that the developed tools and the approaches adopted can be useful to distinguish genomes of closely related organisms of other prokaryotes besides the bacterias considered in this study.

Keywords: Bioinformatics. Genome comparison. Homologous genes. Phylogeny.

Lista de figuras

- Figura 1 – Trecho de um dos arquivos *faa* utilizado no projeto e que contém apenas três sequências. Nesse formato de arquivo, cada sequência inicia-se por uma linha contendo o identificador e a descrição da sequência, seguida pelas linhas contendo os dados da sequência, de modo que cada aminoácido é representado por uma letra 19
- Figura 2 – Exemplo de um alinhamento realizado com quatro sequências de nucleotídeos e envolvendo 55 caracteres, que, nesse caso, é o número de caracteres da sequência de maior comprimento. Em *a.*, as sequências não estão alinhadas e cada uma delas ocupa uma linha do alinhamento. Já em *b.*, as sequências estão alinhadas e esse alinhamento foi realizado de modo a maximizar o número de caracteres idênticos em sequências distintas estivessem dispostos na mesma coluna. As colunas em que isso ocorre estão identificadas por asteriscos (presentes na parte inferior do alinhamento) 25
- Figura 3 – Ilustração dos conceitos de homologia, ortologia e paralogia. Após o gene ancestral, em turquesa, sofrer uma duplicação, uma cópia desse gene, em laranja, também passa a estar presente no genoma e, portanto, ambos os genes são genes parálogos. Caso ocorra a especiação a partir da espécie no qual o gene turquesa estava presente e as novas espécies também apresentem genes que descendam do gene turquesa presente no ancestral em comum, esses genes descendentes existentes nas duas (ou mais) espécies serão ortólogos. Tanto os genes originados pela duplicação quanto os originados pela especiação são também genes homólogos . . . 28

Figura 4 – Exemplo de árvore filogenética na qual são destacados os nós internos, representando os organismos ancestrais, e os ramos, que conectam os indivíduos aos seus ancestrais em comum. Nesse exemplo, os indivíduos representados pelos nós terminais em verde compartilham como ancestral mais próximo em comum o nó interno em laranja, enquanto os nós terminais azuis compartilham como ancestral o nó interno lilás. O nó interno vermelho localizado na raiz da árvore, por sua vez, representa o indivíduo mais ancestral da filogenia que, por meio de processos evolutivos ocorridos ao longo do tempo, originou os nós internos laranja e lilás	32
Figura 5 – Resumo das etapas para se realizar a inferência filogenética a partir de dados genômicos	33
Figura 6 – Filogenia gerada para representar visualmente a árvore filogenética no formato <i>Newick</i> descrita como ((A,B),((C,D),E));	36
Figura 7 – Filogenia gerada para representar visualmente a árvore filogenética no formato <i>Newick</i> descrita como ((A:7.0,B:2.0):3.0,((C:3.0,D:3.0):2.0,E:10):8.0);, na qual se considera o comprimento dos ramos	37
Figura 8 – Distribuição dos artigos encontrados e selecionados em cada base de dados	47
Figura 9 – Distribuição dos artigos encontrados de acordo com o ano de publicação	47
Figura 10 – Distribuição dos artigos encontrados de acordo com a categoria para a qual foram classificados	48
Figura 11 – Resumo das etapas do conjunto de ferramentas para a identificação dos genes homólogos e para a geração das árvores filogenéticas e da rede de genes	65
Figura 12 – Rede de famílias de genes homólogos	74
Figura 13 – Destaque da rede de famílias de genes homólogos. Por meio dela, é possível notar: um componente densamente povoado, contendo diversos genes (A); um componente com baixo coeficiente de agrupamento (B); um componente composto por um único gene de cada um dos 15 genomas (C); e um componente com apenas quatro genes (D)	75
Figura 14 – Filogenia de referência gerada com base nos genes de famílias de genes homólogos	75

Figura 15 – Filogenia gerada com base na região <i>16S rRNA</i> e incluindo o genoma da <i>Xanthomonas albilineans</i> , que foi considerado distante dos demais genomas. Também pode-se observar que não foi possível diferenciar diversos genomas nessa filogenia	77
Figura 16 – Filogenia gerada com base na região <i>16S rRNA</i> sem o genoma da <i>Xanthomonas albilineans</i> . Nessa filogenia também não houve a distinção de alguns genomas	78
Figura 17 – Matriz de distância entre os genomas considerando a presença ou ausência de genes das famílias de genes homólogos, de modo que, quanto maior o número de famílias de genes homólogos compartilhadas pelos genomas, menor é a distância entre eles	79
Figura 18 – Filogenia gerada de acordo com a presença ou ausência de genes nos grupos de homólogos. Nela, é possível observar os três principais grupos de <i>Xanthomonas</i> e que a <i>Xanthomonas albilineans</i> foi colocada próxima das <i>Xanthomonas oryzae</i>	80
Figura 19 – Mapeamento dos genomas de acordo com as duas componentes principais, em que é possível observar os três principais grupos de <i>Xanthomonas</i> e que a <i>Xanthomonas albilineans</i> está localizada entre as <i>Xanthomonas oryzae</i> e <i>Xanthomonas campestris</i>	81
Figura 20 – Filogenia gerada com base nas distâncias calculadas pela ferramenta GGDC utilizando a fórmula 1, na qual há a distinção de todos os genomas e a separação dos três principais grupos de <i>Xanthomonas</i> , além da <i>Xanthomonas albilineans</i> ter sido considerada como distante das demais	82
Figura 21 – Filogenia gerada a partir da utilização da fórmula 1 do GGDC para estimar a probabilidade de que o DDH fosse maior que 70%. Nessa filogenia, é possível observar os três principais grupos de <i>Xanthomonas</i> , mas não foi possível distinguir alguns dos genomas	83
Figura 22 – Filogenia gerada a partir da utilização da fórmula 3 do GGDC para estimar a probabilidade de que o DDH fosse maior que 70%, em que observa-se os três principais grupos de <i>Xanthomonas</i> , mas na qual não foi possível diferenciar nove dos 15 genomas	84

Figura 23 – Filogenia gerada a partir da família de genes <i>30S ribosomal protein S1</i> , considerando o comprimento dos ramos	86
Figura 24 – Filogenia gerada a partir da família de genes <i>30S ribosomal protein S1</i> que não considera o comprimento dos ramos	87
Figura 25 – Filogenia gerada a partir da família de genes <i>50S ribosomal protein L35</i> . A filogenia gerada a partir da família de genes <i>integration host factor subunit alpha</i> apresentou essa mesma topologia. Nelas, apenas a <i>Xanthomonas albilineans</i> foi diferenciada das demais	88
Figura 26 – Filogenia geradas a partir da família de genes <i>acetyl-CoA carboxylase biotin carboxylase subunit</i> , na qual houve uma certa separação dos três principais grupos de <i>Xanthomonas</i> , mas houve a distinção dos diferentes patovares	88
Figura 27 – Filogenia gerada a partir da família de genes <i>50S ribosomal protein L22</i> , na qual houve apenas a separação de dois grupos de <i>Xanthomonas</i>	89
Figura 28 – Filogenia gerada a partir da família de genes <i>50S ribosomal protein L18</i> , na qual nota-se que o grupo de <i>Xanthomonas oryzae</i> não foi propriamente agrupado	90
Figura 29 – Filogenia gerada a partir da família de genes <i>transcription antitermination protein NusG</i> , em que é possível observar que o grupo de <i>Xanthomonas campestris</i> não foi propriamente agrupado	90
Figura 30 – Filogenia mais frequente, gerada a partir das famílias de genes <i>50S ribosomal protein L14</i> e <i>acyl carrier protein</i>	91
Figura 31 – Filogenia gerada em um das iterações utilizando 80% dos genes da <i>Xanthomonas campestris pv. campestris str. 8004</i> , na qual a <i>Xanthomonas campestris pv. campestris str. ATCC 33913</i> está mais próxima evolutivamente de uma <i>Xanthomonas</i> distinta a que estava na árvore de referência	97
Figura 32 – Filogenia gerada em um das iterações utilizando 10% dos genes da <i>Xanthomonas campestris pv. campestris str. 8004</i> , na qual o grupo de <i>Xanthomonas oryzae</i> é o mais distante	99
Figura 33 – Filogenia gerada em um das iterações utilizando 30% dos genes da <i>Xanthomonas oryzae pv. oryzae PXO99A</i> , na qual o grupo de <i>Xanthomonas campestris</i> é o mais distante	100

Lista de tabelas

Tabela 1 – Resumo das informações obtidas nos artigos incluídos na revisão sistemática	50
Tabela 2 – Informações dos genomas utilizados nos experimentos referentes à abreviação do nome do genoma utilizado, seu tamanho, número de genes e a cor com a qual os seus genes foram representados nas redes de genes geradas pela ferramenta de visualização das redes.	72
Tabela 3 – Lista das famílias de genes homólogos presentes em todos os genomas. Em destaque estão as famílias que apresentavam mais de um gene em um genoma e que não foram utilizadas no experimento	85
Tabela 4 – Médias e medianas de correlação das posições e distâncias dos genomas na filogenia quando variada a porção de genes da <i>Xanthomonas albilineans</i>	93
Tabela 5 – Médias e medianas de correlação das posições e distâncias dos genomas na filogenia quando variada a porção de genes da <i>Xanthomonas oryzae pv. oryzae MAFF 311018</i>	94
Tabela 6 – Médias e medianas de correlação das posições e distâncias dos genomas na filogenia quando variada a porção de genes de <i>Xanthomonas oryzae pv. oryzae KACC 10331</i>	94
Tabela 7 – Proporção de genes necessária para que as medianas das correlações de distância e posição fossem acima de 0,7. Entre parênteses estão os valores obtidos nessas medianas e o hífen indica que não foi obtida nenhuma mediana acima de 0,7 para nenhuma proporção testada	95
Tabela 8 – Proporção mínima de número de genes necessária para que as filogenias geradas fossem idênticas à filogenia de referência	97
Tabela 9 – Proporção mínima de genes necessária para que os genomas fossem separados nos três principais grupos de <i>Xanthomonas</i> nas filogenias geradas	99

Sumário

1	Introdução	16
1.1	<i>Objetivos</i>	18
1.2	<i>Metodologia</i>	18
1.3	<i>Organização da dissertação</i>	22
2	Conceitos fundamentais	23
2.1	<i>Sequenciamento e montagem de genomas</i>	23
2.2	<i>Alinhamento de sequências</i>	24
2.3	<i>Anotação de genes</i>	26
2.4	<i>Transferência horizontal de genes</i>	27
2.5	<i>Genes homólogos</i>	27
2.6	<i>Gene e região 16S rRNA</i>	29
2.7	<i>Xanthomonas</i>	30
2.8	<i>Análise filogenética</i>	31
2.9	<i>Formato Newick</i>	35
2.10	<i>Agrupamento hierárquico</i>	35
2.10.1	<i>Complete linkage</i>	37
2.10.2	<i>UPGMA</i>	38
2.11	<i>Ferramentas e conceitos computacionais</i>	38
2.11.1	<i>BLAST</i>	38
2.11.2	<i>GGDC</i>	40
2.11.3	<i>MUSCLE</i>	41
2.11.4	<i>PhyML</i>	43
2.11.5	<i>PHYLIP</i>	43
2.11.6	<i>R</i>	44
3	Trabalhos correlatos	45
3.1	<i>Revisão sistemática</i>	45
3.1.1	<i>I - Grafos</i>	49
3.1.2	<i>II - Árvores</i>	51
3.1.3	<i>III - Aprendizado de máquina</i>	52

3.1.4	IV - Outros métodos	53
3.1.5	Considerações finais sobre a revisão sistemática	54
3.2	<i>Revisão bibliográfica não sistemática</i>	55
3.2.1	Filogenia baseada em sequências	55
3.2.2	Filogenia baseada em características de genomas completos	58
3.2.3	Considerações finais sobre a revisão não sistemática	62
4	Conjunto de ferramentas para comparação de genomas	64
4.1	<i>Identificação de genes homólogos</i>	64
4.2	<i>Comparação dos genomas</i>	67
4.3	<i>Análise e visualização das redes de genes</i>	68
4.4	<i>Ferramentas complementares</i>	69
4.4.1	Seleção de genes em um genoma	69
4.4.2	Comparação de filogenias	70
5	Resultados e discussão	72
5.1	<i>Conjunto de dados</i>	72
5.2	<i>Experimentos</i>	76
5.2.1	Experimento #1 - reconstrução filogenética com base na presença ou ausência de genes das famílias de genes homólogos	79
5.2.2	Experimento #2 - reconstrução filogenética com base na ferramenta GGDC	81
5.2.3	Experimento #3 - reconstrução filogenética com base nos genes ortólogos compartilhadas por todos os genomas	84
5.2.4	Experimento #4 - reconstrução filogenética com base nas proporções de genes dos genomas	91
5.3	<i>Considerações finais</i>	99
6	Conclusões	104
6.1	<i>Principais contribuições</i>	105
6.2	<i>Trabalho futuros</i>	106
	Referências¹	107

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

1 Introdução

A bioinformática, em linhas gerais, consiste no emprego de técnicas computacionais para o estudo e entendimento de problemas e questões biológicas (VERLI, 2014). Apesar dessa definição geral, a bioinformática é um campo complexo e interdisciplinar envolvendo diversas áreas, tais como a biologia, química, física, ciência da computação e estatística.

Nos últimos anos, os avanços tecnológicos ocorridos permitiram que a bioinformática crescesse e se tornasse cada vez mais importante para o progresso dos estudos biológicos, visto que esses avanços provocaram, por exemplo, uma redução dos custos para o sequenciamento de genomas, assim como no tempo necessário para a realização dessa tarefa. Como consequência, o número de genomas sequenciados cresceu drasticamente e, com essa grande quantidade de dados, fez-se necessário o desenvolvimento de ferramentas computacionais para a montagem, armazenamento e análise desses dados (LAND et al., 2015).

Com relação à análise dos genomas, uma das tarefas na qual a bioinformática pode ser aplicada é a comparação de genomas. Essa, por sua vez, é importante para o estudo e entendimento das histórias evolutivas de organismos, permitindo, por exemplo, identificar características que estão presentes em determinados organismos, mas em outros não (MOREIRA, 2015).

Essa comparação é interessante ao ser realizada entre genomas de bactérias, organismos que tiveram o maior número de genomas sequenciados ao longo dos últimos anos. Isso porque o entendimento de suas histórias evolutivas pode auxiliar, por exemplo, na identificação das adaptações moleculares ocorridas em genes e proteínas e que são necessárias para que alguns desses organismos possam sobreviver em condições extremas; na criação de um meio de cultura eficiente que permita que bactérias de difícil cultivo possam crescer em laboratório; e na prevenção do surgimento de novas doenças, assim como no combate às doenças já conhecidas e que são causadas por esses organismos, ao identificar os genes patogênicos existentes em seus genomas (MOREIRA, 2015).

Por muitos anos e até atualmente, uma região específica de RNA conhecida como *16S rRNA* é comumente utilizada para inferir árvores filogenéticas para a visualização das histórias evolutivas de bactérias e de outros procariotos. Esta região é utilizada por ser relativamente bem conservada entre procariotos e por possuir ao menos uma cópia presente em todos os genomas bacterianos, o que permite realizar a comparação entre

esses micro-organismos (LAND et al., 2015). Entretanto, essa região pode apresentar um baixo poder de discriminação filogenética entre organismos de uma mesma espécie ou entre alguns gêneros quando suas sequências são muito similares, pois, muitas vezes, não fornece informações relevantes sobre diferenças e similaridades para organismos muito próximos evolutivamente (JANDA; ABBOTT, 2007).

Foi a importância da comparação de genomas de bactérias e a necessidade de uma metodologia para comparar genomas muito próximos evolutivamente que motivaram o presente trabalho, que visou à implementação de ferramentas para comparar genomas completos e identificar os genes de famílias de genes homólogos presentes em cada um deles. Com base nas famílias de genes homólogos, foram realizados experimentos para analisar se esses genes poderiam ser utilizados para distinguir genomas de organismos muito próximos evolutivamente.

Nos experimentos realizados, as ferramentas foram utilizadas para comparar genomas de bactérias do gênero *Xanthomonas*, que apresenta bactérias que geram grandes prejuízos econômicos por causarem doenças em diversas plantas e cujo estudo é interessante por apresentar grupos de difícil separação do ponto de vista filogenético. Com base nessa comparação e na identificação de genes homólogos, os resultados obtidos foram utilizados para gerar filogenias dos genomas estudados que, por sua vez, foram contrastadas com outras filogenias, dentre as quais uma gerada com base na região do genoma que codifica o *16S rRNA*, de modo que se fosse possível analisar se os genes de famílias de genes homólogos presentes em todos os genomas poderiam ser utilizados para realizar essa inferência filogenética para genomas muito próximos evolutivamente. Os resultados também permitiram analisar as proporções de genes homólogos nos genomas sequenciados que eram necessárias para que fosse possível gerar filogenias desses organismos de modo que os que fossem muito próximos evolutivamente fossem diferenciados (no caso dos sequenciamentos não completos/parciais).

Este trabalho está contextualizado em um projeto maior denominado BIGA¹ (Bioinformática, Genômica e Associados). O BIGA é composto por pesquisadores em biociências e bioinformática e tem como objetivo avançar no estado da arte da biologia computacional por meio do trabalho interdisciplinar em projetos biológicos, que possuem necessidades sofisticadas de bioinformática.

¹ <http://lbi.iq.usp.br/biga>

1.1 Objetivos

O objetivo geral deste trabalho foi propor e implementar ferramentas computacionais para realizar a comparação entre genomas de bactérias, levando-se em consideração seus genes, visando identificar genes de famílias de genes homólogos e, com base em diferentes abordagens para gerar filogenias utilizando esses genes, analisar se os mesmos permitem gerar filogenias nas quais organismos muito próximos evolutivamente estão diferenciados uns dos outros.

Para alcançar este objetivo geral, o projeto possuiu ainda os seguintes objetivos específicos:

- realizar o levantamento do estado da arte na comparação entre genomas de bactérias;
- desenvolver e estender técnicas para a comparação de genomas de bactérias;
- realizar experimentos nos quais as reconstruções filogenéticas fossem baseadas nos genes de famílias de genes homólogos;
- analisar e comparar os resultados obtidos pelos experimentos.

1.2 Metodologia

A metodologia consistiu, primeiramente, no estudo de trabalhos relacionados à comparação de genomas de bactérias e à reconstrução filogenética por meio de revisões bibliográficas. Em seguida, métodos foram escolhidos para serem adotados na especificação e implementação de ferramentas computacionais para a comparação de genomas, assim como também foram escolhidas ferramentas desenvolvidas por outros autores para serem utilizadas no trabalho.

Para realizar os experimentos, foram utilizados 15 genomas completos de bactérias do gênero *Xanthomonas*. Esses genomas foram obtidos no banco de dados biológicos do NCBI (*National Center for Biotechnology Information*), o GenBank², no dia 25 de julho de 2015 e estavam representados por sequências de aminoácidos em arquivos de texto no formato FASTA *faa* (*FASTA amino acid*). A figura 1 apresenta um trecho de um arquivo FASTA utilizado no presente projeto e que serviu como entrada para a ferramenta BLASTP³ que, por sua vez, foi utilizada para realizar o alinhamento das sequências.

² <http://www.ncbi.nlm.nih.gov/genbank>

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

Figura 1 – Trecho de um dos arquivos *faa* utilizado no projeto e que contém apenas três sequências. Nesse formato de arquivo, cada sequência inicia-se por uma linha contendo o identificador e a descrição da sequência, seguida pelas linhas contendo os dados da sequência, de modo que cada aminoácido é representado por uma letra

```
>gi|285016822|ref|YP_003374533.1| chromosomal replication initiator protein DnaA [Xanthomonas albilineans GPE PC73]
MDAWPRCLERLEAEFPAEDVHTWLPKPLQAEQRHDSMVLYAPNAFIVEQVHARYLSRIRELLNYFAGSED
VLLQVGRSPRAPEPMAAASDSTTARTPSPQLLEPFAGNLDHYTFANFVEGRSNQLGLAAAVQAAQKPGD
RAHNPLLLYGSTGLKTHLMFAAGNAMREQNPNAVYMLRSEQFFSAMIRALQEKTMDFKRQFQQVDAL
LIDDIQFFAGKDRTOEEFFHTFNALFDGRQOIILTCDRYPREVEGLEPRLSRLANGLSVAIDPPDFETR
AAIVLAKARERGAEPDEVAFLIAKKMRSNVRDLEGALNTLAARANFTGRAITTEFAQETLRDLLRAQQQ
AIGIPNIQKTVADYYGLQIKDLLSKRRTRSLARPRQMAMALTKELTEHSLEIGDAFAGRDHTTVLHACR
QIKHLMETDGKLRDNDKLRKLSLSE
>gi|285016823|ref|YP_003374534.1| DNA polymerase III subunit beta [Xanthomonas albilineans GPE PC73]
MRFTLQREAFLLKPLAQVNVVVERRQTLPLVLANFLVQVHGGQLSLTGTDLLEVHVSRIAVDDAQDGETTIP
ARKLFEIVRALPDGSKVTVSLSDKITVQTGRSFTLATLPANDFPSVDEVETTERVVQEAALKELIER
TAFAMAQQDVRYVYLLNGLLFDLHDKTLRCVATDGHRLALCETELNAGGAKRQIIVPRKGVTELQRLLEGG
DREVELEVGRSHVRVKNRNDVFTSKLIDGRFPDYETVIPIGADREVKLDREILRAALQRAAAILSNKEYRG
VRVEVSPGQLKISAHNPQEQAEEVEADTKVSDLAIGFNVNLYLLDALSALREEFVVIQLRDNASSALVR
EASSARSRHVMPRLRL
>gi|285016824|ref|YP_003374535.1| DNA replication and repair protein RecF [Xanthomonas albilineans GPE PC73]
MHLFRDLFHHLLRRFSPIELTPAPGMNLITGDNAGAKTTILEAMHLMAYGRSFRRCVRDGLIRQGGAGLDV
FVEWHQAADTGQTRRRRAGLQHSQGQDWRGRLDGQDVPLYGLTCAALAVVTFEPGSHVLVSGGGEARRRF
VDWGLFHVPEPDFLSLWRRYARALKQRNALLKSGGAGAALDAWDHLEAGEPLTSRRQHYLDRLRLRLLI
LAPELAPALGIEQLQFSPGWRREISLADALLNDRDRQLGYTTVGPHRADWSLSFAAIPGRDALSRGQ
AKLTALACLLAQAEYAEQRGEPVIALDDLASELDQHHQARVLQRLRAGPAQVFLTATETPQALSDAGL
PIARFHVHGGILSSS
```

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

Com base nesses alinhamentos, foram utilizadas ferramentas desenvolvidas em Java⁴ para a comparação dos genomas e identificação das famílias de genes homólogos presentes nesses genomas.

Ao todo, quatro experimentos foram realizados com diferentes abordagens para realizar a reconstrução filogenética dos 15 genomas de *Xanthomonas* a partir dos dados referentes às famílias de genes homólogos identificadas, exceto no segundo experimento, no qual foram utilizados os genomas completos para gerar as filogenias.

No primeiro experimento, a tabela com dados binários sobre a ausência ou presença de genes de cada família de genes homólogos em cada genoma foi utilizada para geração de uma matriz de distância calculada com base na distância euclidiana. Em seguida, utilizou-se a função *hclust* presente no *software* R⁵ para realizar um agrupamento hierárquico com o método *complete linkage*, que é o padrão da função, a partir da matriz de distância para, com isso, gerar a árvore filogenética. Importante salientar o método *complete linkage* foi utilizado nos experimentos porque constatou-se que as árvores geradas com os outros métodos disponíveis na função *hclust* eram similares e, portanto, optou-se por utilizar o método padrão da função.

⁴ <http://www.java.com>

⁵ <http://www.r-project.org>

No segundo experimento, por sua vez, utilizou-se a ferramenta GGDC⁶ para calcular a distância entre os genomas. Além da distância, a ferramenta também fornece um valor estimado de DDH (hibridização DNA-DNA) e a probabilidade de que o DDH seja superior a 70%. Para esses dois últimos valores, que são porcentagens, as medidas de distância entre os genomas foram dadas ao subtraí-los do número 100. Os três valores fornecidos pela ferramenta foram utilizados separadamente para gerar as filogenias dos genomas, de modo que as distâncias entre os genomas foram inseridas em matrizes de distância, que serviram como entrada para a ferramenta PHYLIP⁷. Nessa ferramenta, as filogenias foram geradas utilizando o método UPGMA presente no pacote *neighbor*, que são retornadas no formato *Newick*. Por fim, os dendrogramas para representação visual das filogenias foram gerados com o *software* R após a leitura da árvore com a função *read.tree*.

Já o terceiro experimento utilizou a mesma tabela do primeiro experimento para identificar e selecionar os genes das famílias de homólogos presentes em todos os genomas para, com cada um deles, realizar a reconstrução filogenética. Após a identificação dos genes homólogos compartilhados por todos os genomas, foram utilizadas duas ferramentas: o MUSCLE⁸ e o PhyML⁹ para, respectivamente, realizar o alinhamento múltiplo dos genes de todos os genomas e gerar as árvores filogenéticas a partir dos alinhamentos múltiplos. Além disso, como o PhyML retorna a filogenia no formato *Newick*, também foi utilizada a função *read.tree* do pacote *ape* do *software* R para gerar a representação visual das árvores filogenéticas.

Por fim, para o quarto experimento foi desenvolvido um algoritmo, também em Java, para selecionar, aleatoriamente, genes de famílias de genes homólogos presentes no genoma testado até alcançar a proporção de genes delimitada para o teste. Em seguida, matrizes de distância entre os genomas foram calculadas considerando a distância euclidiana. Por fim, árvores filogenéticas foram construídas a partir das matrizes de distância utilizando a função *hclust* do *software* R, que realizou o agrupamento hierárquico com o método *complete linkage*. Para o experimento realizado, as porcentagens de genes selecionados para cada um dos 15 genomas variaram de 10 a 90% e, para cada proporção, o experimento foi realizado 10 vezes, visto que os genes eram selecionados aleatoriamente. Com os resultados obtidos após as 10 iterações, calculou-se a média de correlação da posição dos genomas na

⁶ <http://ggdc.dsmz.de>

⁷ <http://evolution.genetics.washington.edu/phylip>

⁸ <http://www.ebi.ac.uk/Tools/msa/muscle>

⁹ <http://www.atgc-montpellier.fr/phyml>

árvore filogenética e da correlação de suas distâncias para os demais genomas com relação às distâncias e posições na filogenia de referência.

Com relação à árvore filogenética de referência utilizada para ser comparada com todas as filogenias resultantes dos quatro experimentos, usou-se a matriz de distância calculada com base na distância euclidiana a partir da tabela contendo o número de genes de cada família de genes homólogos presentes em cada um dos genomas estudados. A partir dessa matriz de distância, utilizou-se novamente a função *hclust* do *software* R para realizar o agrupamento hierárquico com base no método *complete linkage* e, com isso, gerar a árvore de referência.

Além disso, como as filogenias geradas no terceiro experimento estavam no formato *Newick*, a filogenia de referência também foi convertida para esse formato. As topologias das árvores filogenéticas do terceiro experimento foram então comparadas com a da árvore filogenética de referência por meio de um algoritmo desenvolvido, também utilizando a linguagem de programação Java, para realizar essa comparação automaticamente.

Adicionalmente, a ferramenta PhyML também foi utilizada para gerar a filogenia a partir dos resultados de alinhamentos múltiplos realizados com base na região *16S rRNA*, amplamente utilizada como referência para a reconstrução filogenética, para que ela também fosse comparada com as filogenias geradas pelos experimentos.

Finalmente, a avaliação dos experimentos foi realizada com a comparação dos resultados obtidos nos quatro experimentos, tanto comparando suas filogenias umas com as outras, quanto com a filogenia de referência e com a gerada a partir da região *16S*. Além disso, especialistas do domínio também fizeram a validação das ferramentas e dos resultados obtidos.

É importante salientar que esta seção descreve as ferramentas desenvolvidas neste projeto e os experimentos realizados em linhas gerais. Descrições detalhadas sobre esses tópicos podem ser obtidas, respectivamente, nos capítulos 4 e 5. Além disso, detalhes sobre as ferramentas elaboradas por outros autores e que foram utilizadas neste projeto podem ser encontrados na seção 2.11.

1.3 Organização da dissertação

Esta dissertação está organizada em seis capítulos, incluindo o desta introdução. Os demais capítulos estão organizados de modo que:

- o capítulo 2 aborda os fundamentos básicos e necessários para o entendimento do trabalho apresentado nesta dissertação;
- o capítulo 3 compreende os trabalhos relacionados ao presente projeto e que foram encontrados a partir de revisões bibliográficas;
- o capítulo 4 descreve o conjunto de ferramentas desenvolvido e utilizado para a identificação de genes homólogos e para a geração das árvores filogenéticas e da rede de genes;
- o capítulo 5 apresenta e discute sobre os resultados obtidos nos experimentos realizados;
- e, por fim, o capítulo 6 traz as conclusões do trabalho, apresentando suas principais contribuições, bem como os trabalhos futuros esperados.

2 Conceitos fundamentais

Este capítulo aborda conceitos fundamentais com o objetivo de proporcionar um melhor entendimento do projeto descrito nesta dissertação. Esses conceitos, apresentados nas próximas seções, abrangem tanto os computacionais, biológicos e de bioinformática quanto os que dizem respeito às ferramentas computacionais que foram utilizadas no projeto.

2.1 Sequenciamento e montagem de genomas

Para que o estudo de genomas seja possível é necessário, primeiramente, realizar o sequenciamento e a montagem dessas sequências genômicas.

O sequenciamento de genomas consiste em uma técnica para identificar a ordem correta da sequência de nucleotídeos de uma molécula de DNA ou de RNA, objetivando-se conhecer a informação genética presente nessas moléculas. Esse sequenciamento e uma boa montagem das sequências permitem obter informações interessantes para o estudo de genomas, tais como a estrutura e função dos genes; presença de genes adquiridos por transferência horizontal; e as relações evolutivas entre os organismos.

Os avanços tecnológicos ocorridos nos últimos anos proporcionaram melhorias nesse processo de sequenciamento, visto que desde o surgimento das primeiras metodologias desenvolvidas no final da década de 70, cujo sequenciamento era realizado manualmente e para poucas kilobases, até as tecnologias atuais, chamadas de “sequenciamento de nova geração”, tornou-se possível sequenciar grandes quantidades de sequências em paralelo e em um curto período de tempo. Apesar desses avanços, os atuais sequenciadores ainda possuem uma limitação com relação ao tamanho dos fragmentos de DNA a serem sequenciados, uma vez que, de forma geral, eles ainda não possuem a capacidade de sequenciar longos fragmentos. Desse modo, é necessário isolar e sequenciar apenas fragmentos de interesse ou fragmentar moléculas grandes de DNA para posteriormente fazer o seu sequenciamento. Para esse último caso, faz-se necessário realizar a montagem dos fragmentos da sequência genômica obtidos a partir do sequenciamento com o intuito de tentar obter a sequência original.

Um dos grandes desafios no processo de montagem é colocar os milhares de fragmentos resultantes na ordem correta em que se encontravam no genoma. Para isso, foram desenvolvidos programas para realizar o alinhamento dos fragmentos gerados, chamados de *reads*, baseando-se em suas regiões de sobreposição, o que resulta em sequências únicas chamadas de *contigs*. Desse modo, espera-se que, ao final do processo de montagem, obtenha-se uma sequência única para genomas que possuem um único cromossomo ou várias sequências contíguas que representam o número total de cromossomos do organismo em questão.

No entanto, antes de realizar a montagem, é necessário refinar os *reads* resultantes do sequenciamento, visto que alguns dos *reads* podem conter erros, bem como podem existir trechos de sequências não pertencentes ao organismo sequenciado. Outro problema também existente no processo é a montagem de regiões repetidas do genoma, visto que há dificuldade em realizar a ordenação correta dessas regiões.

Além disso, apesar do desejo de se obter uma sequência única para representar o genoma completo, a maioria dos genomas sequenciados e montados atualmente estão incompletos e ainda não foram finalizados. Isso ocorre por diversos motivos, por exemplo, devido aos erros causados pelo sequenciamento, baixa cobertura proporcionada pelos *reads* sequenciados, genomas com muitas regiões repetitivas, sequenciamento de muitos *reads* com baixa qualidade, contaminação do material genético gerando *reads* que não pertencem ao organismo de interesse, dificuldade em sequenciar regiões complexas do genoma, etc. (MOREIRA, 2015).

2.2 Alinhamento de sequências

O alinhamento é uma técnica computacional complexa na qual se pretende representar uma hipótese evolutiva das sequências, visto que se pode assumir que sequências distintas que apresentam um determinado grau de similaridade quando alinhadas compartilharam um ancestral em comum no passado (e, portanto, estão relacionadas evolutivamente) e que, após a separação do ancestral em comum, cada sequência acumulou diferentes variações ao longo da evolução.

O alinhamento de sequências tem como objetivo estabelecer a melhor correspondência entre as posições/caracteres (que integram as colunas do alinhamento) de

duas ou mais sequências de nucleotídeos ou de aminoácidos (organizadas nas linhas do alinhamento), sem, no entanto, alterar a ordem dos caracteres presentes nessas sequências. Para tanto, a comparação das sequências é realizada comparando os caracteres que representam os nucleotídeos ou aminoácidos. As lacunas (em inglês, *gaps*), que podem ser inseridas nos alinhamentos para que todas as sequências tenham o mesmo comprimento e também para que seja possível visualizar a similaridade entre as sequências com maior facilidade, correspondem a um ou mais eventos de inserção ou deleção de nucleotídeos ou aminoácidos em determinadas posições das sequências e costumam ser representadas, nos alinhamentos, por um hífen ou por um ponto. Os algoritmos para alinhamento visam a minimizar as diferenças entre as sequências e a similaridade entre elas é computada em termos de identidade, que calcula a porcentagem de caracteres idênticos existentes nas mesmas posições em sequências diferentes. A figura 2 apresenta um exemplo de um alinhamento de sequências de nucleotídeos.

Figura 2 – Exemplo de um alinhamento realizado com quatro sequências de nucleotídeos e envolvendo 55 caracteres, que, nesse caso, é o número de caracteres da sequência de maior comprimento. Em *a.*, as sequências não estão alinhadas e cada uma delas ocupa uma linha do alinhamento. Já em *b.*, as sequências estão alinhadas e esse alinhamento foi realizado de modo a maximizar o número de caracteres idênticos em sequências distintas estivessem dispostos na mesma coluna. As colunas em que isso ocorre estão identificadas por asteriscos (presentes na parte inferior do alinhamento)

Fonte: Verli (2014)

Os alinhamentos podem ser realizados de diferentes maneiras. Com relação ao número de sequências a serem alinhadas, há o chamado alinhamento par-a-par (do inglês, *pairwise alignment*), em que o alinhamento é realizado apenas entre duas sequências, e o

alinhamento múltiplo (do inglês, *multiple sequence alignment*), em que o alinhamento é realizado entre três ou mais sequências. Quanto mais sequências são comparadas, maior é o tempo necessário para a realização do alinhamento e, dependendo de quais sequências que serão alinhadas, maior é a dificuldade para que os algoritmos obtenham um alinhamento ótimo.

Já com relação à estratégia utilizada para alinhar as sequências, os alinhamentos podem ser classificados, basicamente, em globais ou locais. No alinhamento global, todos os caracteres que representam os nucleotídeos ou aminoácidos de todas as sequências são alinhados uns com os outros na extensão completa da sequência de maior comprimento, que é o caso do alinhamento realizado na figura 2. Já no alinhamento local, apenas as regiões que apresentarem a mais alta densidade de identidade são alinhadas e, desse modo, quanto maior o grau de similaridade entre as sequências, maior é a extensão do alinhamento local.

Dentre as diversas aplicações em que a análise dos alinhamentos pode ser empregada, pode-se citar a análise de padrões nas sequências, identificação de grupos de sequências relacionadas, montagem de genes e genomas, caracterização de inserções e deleções ocorridas nas sequências e os estudos de inferência estrutural e funcional de proteínas ([VERLI, 2014](#); [MOREIRA, 2015](#)).

2.3 Anotação de genes

Após realizar o sequenciamento e a montagem do genoma, outras duas etapas podem ser realizadas: a predição e anotação de genes.

Para a predição gênica, são utilizados programas com o intuito de buscar e identificar sequências de nucleotídeos correspondentes a cada um dos genes ou de outras regiões de interesse existentes no genoma sequenciado. Para tanto, esses programas reconhecem nucleotídeos que são característicos desses genes ou das regiões de interesse.

Já o processo de anotação trata de identificar os diversos elementos presentes no genoma sequenciado e atribuir, a cada um deles, o máximo de informações biológicas possível, tal como a identificação dos genes e a atribuição de suas respectivas funções com o intuito de se entender os fenômenos biológicos que ocorrem no organismo em questão. Essa anotação pode ser realizada automaticamente ou manualmente, de modo que a primeira é realizada por meio de programas de bioinformática capazes de anotar o conjunto gênico

de um organismo de uma só vez, enquanto a segunda é muito mais demorada, pois é efetuada por um anotador para cada gene separadamente, mas também é feita com mais cuidado e é considerada mais confiável do que a anotação automática. Por conta do tempo gasto para se realizar a anotação manual e por conta do grande volume de dados gerados atualmente, a anotação automática é mais comumente utilizada. É interessante ressaltar que parte das sequências depositadas nos bancos de dados públicos está erroneamente anotada e, mesmo assim, essas sequências servem de base para a anotação gênica. Desse modo, a propagação do erro constitui no maior problema que a tarefa de anotação enfrenta atualmente ([MOREIRA, 2015](#)).

2.4 Transferência horizontal de genes

A transferência horizontal (ou lateral) de genes é um processo no qual partes de DNA podem ser transferidas entre organismos com diferentes constituições genéticas (por exemplo, de diferentes espécies). Esse processo contrasta com a usual transferência vertical, na qual a informação genética é passada apenas pela divisão celular ([ZAHA; FERREIRA; PASSAGLIA, 2003](#)).

Esse processo, cujas ocorrências são observadas principalmente em procariotos, tem importantes implicações na evolução de organismos, já que ele possibilita que, por exemplo, uma determinada bactéria adquira um gene que lhe forneça resistência à antibióticos ([OCHMAN; LAWRENCE; GROISMAN, 2000](#)).

Apesar das vantagens evolutivas que a transferência horizontal de genes pode fornecer aos organismos, ela também pode dificultar as inferências filogenéticas, principalmente aquelas baseadas em um único gene, visto que se houver a transferência horizontal desse gene de referência entre organismos distantes evolutivamente, a árvore filogenética pode acabar por deixá-los próximos um do outro.

2.5 Genes homólogos

Genes homólogos são dois ou mais genes que derivam de um mesmo gene ancestral em comum. Os genes homólogos podem ainda ser divididos em genes ortólogos e genes parálogos.

Genes ortólogos são aqueles gerados por meio do processo de especiação de forma que esses genes, presentes em espécies diferentes, apresentam sequências de nucleotídeos similares e que codificam proteínas similares com funções correspondentes, descendendo de um mesmo gene presente no último ancestral em comum dessas espécies.

Já os genes parálogos são aqueles originados por meio de um evento de duplicação, em que um gene passa a ter mais de uma cópia em um genoma, de modo que, após essa duplicação, as cópias podem divergir do gene original a partir de mutações ocorridas durante o processo evolutivo e podem adquirir funções relacionadas, mas distintas, à original (ZAHA; FERREIRA; PASSAGLIA, 2003). A figura 3 ilustra esses conceitos.

Figura 3 – Ilustração dos conceitos de homologia, ortologia e paralogia. Após o gene ancestral, em turquesa, sofrer uma duplicação, uma cópia desse gene, em laranja, também passa a estar presente no genoma e, portanto, ambos os genes são genes parálogos. Caso ocorra a especiação a partir da espécie no qual o gene turquesa estava presente e as novas espécies também apresentem genes que descendam do gene turquesa presente no ancestral em comum, esses genes descendentes existentes nas duas (ou mais) espécies serão ortólogos. Tanto os genes originados pela duplicação quanto os originados pela especiação são também genes homólogos

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

O número de genes homólogos de uma família de genes pode variar entre os diferentes organismos em decorrência de ganhos, perdas ou duplicações gênicas ocorridas após o processo de especiação (MOREIRA, 2015).

2.6 Gene e região 16S rRNA

Após pesquisas realizadas a partir da década de 80, constatou-se que trechos estáveis dos genomas poderiam ser comparados para se identificar relações filogenéticas entre organismos diferentes (WOESE, 1987). No caso das bactérias, o trecho mais utilizado atualmente para realizar essa comparação é o do gene que codifica o RNA ribossomal 16S (16S rRNA), comumente chamado de *gene 16S rRNA*, que permite não só a comparação entre bactérias como também com os genes 16S rRNA e 18S rRNA existentes, respectivamente, em arqueobactérias e eucariotos (CLARRIDGE, 2004). Além disso, esse gene é muito utilizado por estar presentes em todas as bactérias e por apresentar um alto grau de conservação em sua funcionalidade, permitindo a comparação filogenética (WOESE, 1987).

A região do genoma que codifica o 16S rRNA possui cerca de 1.550 pares de bases e apresenta regiões conservadas, para todos os gêneros de bactérias, e variáveis que evoluem em diferentes taxas, de modo que a comparação filogenética pode ser realizada ao estimar as distâncias entre os genomas a partir dessas regiões variáveis ou levando-se em conta toda a região 16S rRNA (HAUBEN et al., 1997; CLARRIDGE, 2004). Um estudo realizado com genomas de bactérias do gênero *Xanthomonas*, por exemplo, indicou que cerca de 94,34% da sequência do gene 16S rRNA é conservada, 3,99% é variável e 1,66% consiste de posições ambíguas ou lacunas (HAUBEN et al., 1997).

A utilização do gene 16S rRNA, no entanto, apresenta algumas limitações. Isso porque há bactérias que podem apresentar mais de uma cópia desse gene em seu genoma, podendo chegar a 15 ou mais cópias, sendo que algumas delas podem ter sido adquiridas de outros organismos por meio da transferência horizontal. Muitas vezes, essas cópias apresentam sequências diferentes entre espécies diferentes ou até mesmo entre as que estão presentes em um mesmo genoma (VĚTROVSKÝ; BALDRAN, 2013).

Além disso, o 16S rRNA não permite a diferenciação entre espécies e até mesmo alguns gêneros que apresentem sequências idênticas ou muito similares (CLARRIDGE, 2004; JANDA; ABBOTT, 2007). Desse modo, esse gene não é adequado para ser utilizado quando, por exemplo, há o intuito de realizar a diferenciação entre espécies muito próximas de um determinado gênero, visto que as sequências desse gene serão muito similares para organismos muito próximos evolutivamente, ou quando se deseja comparar cepas de uma

mesma espécie, visto que provavelmente não haverá variação significativa das sequências (CLARRIDGE, 2004). Por exemplo, no estudo já citado anteriormente, a utilização do gene *16S rRNA* não foi suficiente para distinguir algumas cepas de *Xanthomonas*, visto que não foram encontradas diferenças nas sequências desse gene (HAUBEN et al., 1997).

2.7 *Xanthomonas*

O gênero *Xanthomonas* compreende um grande grupo de bactérias fitopatógenas, ou seja, são bactérias que causam doenças em plantas. Essas bactérias podem adentrar diversas partes de diferentes espécies de plantas, tais como as folhas, caules, sementes, galhos e frutos, e que podem apresentar sintomas. Esses sintomas podem variar desde lesões e manchas em folhas até surgimento de cancrios e a ocorrência de murchidão, podridão, hipertrofia (excesso de crescimento celular e de tecidos) e hiperplasia (excesso de multiplicação celular) (RUDOLPH, 1993). Muitas das infecções causadas por essas bactérias acarretam danos severos em frutas e vegetais, o que resulta, conseqüentemente, em perdas e prejuízos econômicos. Por esse motivo, estudos relacionados a essas bactérias são importantes, visto que um melhor entendimento de suas patogenicidades e histórias evolutivas pode ajudar no combate de doenças já existentes e a prevenir o surgimento de novas doenças (ZHANG et al., 2015).

Desde o sequenciamento dos dois primeiros genomas de *Xanthomonas* em 2002, muitos outros genomas dessas bactérias foram sequenciados ao longo dos anos. Atualmente, o GenBank contém aproximadamente 400 sequências genômicas de *Xanthomonas*, das quais 50 foram completamente sequenciadas. Por outro lado, grande parte dos outros genomas sequenciados que foram gerados por sequenciadores de alto desempenho consiste de fragmentos com qualidade variada e com propagação de anotações erroneamente feitas. Desse modo, tem-se acesso a uma grande variedade de genomas de bactérias do gênero *Xanthomonas*, tanto de diferentes espécies quanto de patovares, que variam com relação à qualidade de sua montagem e anotação gênica (JACQUES et al., 2016).

2.8 Análise filogenética

A filogenética é uma área pertencente à biologia evolutiva que visa à reconstrução das relações evolutivas entre organismos, tanto dos organismos existentes atualmente quanto dos que já foram extintos, considerando que eles compartilham uma origem evolutiva em comum. Para representar essa reconstrução, são comumente utilizadas árvores filogenéticas, também chamadas de filogenias, que permitem observar as diferentes relações entre os organismos e seus possíveis ancestrais em comum.

Uma árvore filogenética consiste em um diagrama ramificado no qual os ramos conectam os nós terminais aos seus ancestrais, representados pelos nós internos da árvore, de modo que cada nó terminal representa um indivíduo. A figura 4 ilustra os componentes de uma árvore filogenética.

Os indivíduos representados na árvore podem ser, por exemplo, espécies diferentes ou organismos ainda mais próximos evolutivamente como subespécies, que são subdivisões de uma espécie que foram diferenciadas com base em suas variações fenotípicas ou por formarem diferentes grupos de cepas distinguidos por suas variações genéticas, ou patovares, que são subdivisões de espécies ou subespécies de bactérias que se diferenciam dos demais de acordo com as plantas que podem infectar (BRENNER; STALEY; KRIEG, 2005; SCHAECHTER, 2009).

Os nós terminais são também chamados de unidades taxonômicas operacionais (do inglês, *operational taxonomical units* - OTUs) e a ordem e a disposição exata desses nós em uma filogenia é denominada topologia. Cada nó interno da árvore representa um indivíduo ancestral que dá origem a dois ou mais ramos. Do ponto de vista computacional, as árvores filogenéticas podem ser entendidas como um agrupamento hierárquico representado por um dendrograma.

Para gerar as árvores filogenéticas é necessário, primeiramente, selecionar os dados dos genomas que serão utilizados para a reconstrução filogenética. De modo geral, após a obtenção dos dados genômicos a partir do sequenciamento de DNA, conjunto de genes ortólogos são agrupados em famílias de genes objetivando-se a análise filogenética. Essa avaliação de homologias ou ortologias é uma etapa crucial e é quase sempre baseada na comparação de similaridade. Muitos dos métodos utilizados para a reconstrução filogenética

Figura 4 – Exemplo de árvore filogenética na qual são destacados os nós internos, representando os organismos ancestrais, e os ramos, que conectam os indivíduos aos seus ancestrais em comum. Nesse exemplo, os indivíduos representados pelos nós terminais em verde compartilham como ancestral mais próximo em comum o nó interno em laranja, enquanto os nós terminais azuis compartilham como ancestral o nó interno lilás. O nó interno vermelho localizado na raiz da árvore, por sua vez, representa o indivíduo mais ancestral da filogenia que, por meio de processos evolutivos ocorridos ao longo do tempo, originou os nós internos laranja e lilás

Fonte: Verli (2014)

são baseados ou em sequências ou em características de genomas completos. Essas etapas para a inferência filogenética podem ser visualizados na figura 5.

Nas abordagens que geram filogenias baseando-se em sequências, é necessário, primeiramente, que genes ortólogos sejam alinhados por meio de ferramentas de alinhamento múltiplo de sequências. Após esse alinhamento, duas abordagens distintas podem ser utilizadas para inferir as árvores filogenéticas: supermatrizes e superárvores. Na abordagem de supermatrizes, é analisada a concatenação dos genes individuais, sendo que os métodos de verossimilhança para a reconstrução filogenética são mais adequados para a análise dessas supermatrizes. Já na abordagem de superárvores, as árvores ótimas obtidas a partir da análise individual dos genes são combinadas para se obter a filogenia.

As abordagens que geram filogenias com base nas características dos genomas, por sua vez, podem levar em consideração o conteúdo gênico ou a ordem dos genes. As abordagens que consideram o conteúdo gênico dos genomas reconstroem as filogenias com base em distâncias que representam a proporção de genes ortólogos compartilhados pelos

Figura 5 – Resumo das etapas para se realizar a inferência filogenética a partir de dados genômicos

Fonte: adaptado de Delsuc, Brinkmann e Philippe (2005)

genomas. Tais distâncias podem ser calculadas com base em algoritmos de distância ou a partir de matrizes referentes à presença ou ausência de genes homólogos ou ortólogos nos genomas usando o método de máxima parcimônia. Métodos que consideram a ordem de genes, por sua vez, reconstroem as filogenias ao minimizar o número de *breakpoints*, que são posições em que podem ter ocorrido deleções, inversões ou permutações em genomas, ou ao computar a presença ou ausência de pares de genes ortólogos (DELSUC; BRINKMANN; PHILIPPE, 2005).

Após a identificação de quais e como os dados genômicos serão considerados para a reconstrução das árvores filogenéticas, métodos baseados em distância ou caracteres podem ser utilizados para comparar e computar a similaridade entre os genomas.

Os métodos de distância fornecem uma medida de distância, por exemplo, entre cada par de sequências de um alinhamento, de modo que, a partir das medidas calculadas, uma matriz de distância é gerada e, a partir dessa matriz, a filogenia é construída. Entretanto, apesar da simplicidade e rapidez para se gerar árvores filogenéticas a partir de matrizes de distância, esse método pode ser pouco realista, visto que se perde informação ao converter os caracteres das sequências para medidas de distância.

Já os métodos baseados em caracteres utilizam os caracteres que representam os nucleotídeos ou aminoácidos nas sequências alinhadas. Dentre esses métodos, pode-se citar o de máxima parcimônia e o de máxima verossimilhança.

Para o método de máxima parcimônia, a melhor hipótese evolutiva é aquela em que é necessário o menor número de passos para explicar os dados, ou seja, a árvore filogenética ideal é a que possuir o menor número de mudanças para explicar os dados dos alinhamentos realizados. Nesse método, não há cálculo da distância e as árvores são calculadas a partir dos dados dos alinhamentos. As possíveis árvores geradas são comparadas e cada uma delas recebe uma pontuação, que indica o número mínimo de mudanças dos nucleotídeos ou aminoácidos necessárias ao longo da evolução para posicionar as sequências na árvore. O método de máxima parcimônia é relativamente rápido e apresenta um bom desempenho para sequências altamente similares, mas apresenta um baixo desempenho quando há divergências consideráveis entre as sequências.

Já o método de máxima verossimilhança tem como objetivo encontrar a árvore filogenética que maximiza a probabilidade dos dados alinhados. Para isso, são calculadas as probabilidades de cada uma das topologias geradas e a árvore considerada ótima é aquela que apresentar o maior valor de verossimilhança (maior probabilidade dos resultados observados terem sido gerados de acordo com o modelo de substituição dos nucleotídeos ou aminoácidos). Apesar desse método ser considerado robusto e consistente, apresenta um alto custo computacional e é sensível a ambiguidades presentes nos alinhamentos.

Após a reconstrução da árvore, é necessário avaliar o grau de confiança da topologia obtida. Essa avaliação pode ser realizada utilizando uma abordagem denominada *bootstrapping*, em que as colunas do alinhamento original são reamostradas e novas árvores são geradas a partir desses novos conjuntos de sequências. As árvores geradas são então comparadas e, para cada nó da árvore, é associado um valor de porcentagem indicando o quão frequentemente as duas sequências ligadas pelo nó ocorrem juntas em diferentes árvores, de modo que, quanto maior for o valor de *bootstrap*, maior é a probabilidade de que as sequências foram posicionadas corretamente na filogenia (VERLI, 2014; MOREIRA, 2015).

2.9 Formato Newick

O *Newick* é um formato definido informalmente e adotado por diversas ferramentas computacionais de inferência filogenética para representar árvores filogenéticas sob a forma de texto. Nele, os nós internos e grupos descendentes de cada nó interno que estão presentes na árvore são distinguidos uns dos outros por meio da utilização de parênteses, os nós terminais de um mesmo grupo são separados uns dos outros por meio de vírgulas e a descrição da árvore termina com um ponto e vírgula.

Como exemplo desse formato, podemos observar a filogenia dada pela figura 6, cuja representação no formato *Newick* seria $((A,B),((C,D),E));$, de modo que os parênteses separam os dois grupos principais descendentes de dois nós internos diferentes, um composto pelos nós terminais *A* e *B* e outro por *C*, *D* e *E*, sendo que, neste último grupo, os nós terminais *C* e *D*, que estão mais próximos evolutivamente, são separados de *E* por meio de parênteses.

Também é possível adicionar o comprimento dos ramos nesse formato, representado por um número real. Para tanto, o comprimento é acrescentado após a definição de um nó, seja ele interno ou terminal, e é precedido pelo símbolo de dois pontos. Adicionando distâncias de exemplo na árvore anterior, de modo que ela seja representada no formato *Newick* como $((A:7.0,B:2.0):3.0,((C:3.0,D:3.0):2.0,E:10):8.0);$, obtêm-se a árvore presente na figura 7.

2.10 Agrupamento hierárquico

O agrupamento hierárquico é um método para realização de agrupamentos, que, por sua vez, consiste em um processo no qual objetos (no caso deste trabalho, os objetos são os genomas) são associados a grupos de objetos similares com base em suas características (no caso deste trabalho, são os genes de famílias de genes homólogos). Deste modo, um grupo é uma coleção de objetos que são similares uns com os outros, mas que são dissimilares à objetos de outros grupos. O processo de agrupamento tem inúmeras aplicações, tais como nas áreas de reconhecimento de padrões, mineração de dados e em bioinformática.

No agrupamento hierárquico, os objetos são agrupados em uma árvore de grupos, de modo que os objetos são representados pelos nós terminais da árvore e os nós internos

Figura 6 – Filogenia gerada para representar visualmente a árvore filogenética no formato *Newick* descrita como $((A,B),((C,D),E))$;

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

indicam a organização desses objetos de acordo com suas similaridades. Essa árvore é comumente representada por meio de um diagrama chamado de dendrograma, que permite observar como os grupos foram se formando.

A similaridade pode ser calculada para pares de objetos ou pares de grupos utilizando uma medida de distância, tal que, quanto menor for o valor da distância, mais similares os objetos ou os grupos são. Logo, os objetos ou grupos com a menor distância entre eles são agrupados primeiro.

Uma medida de distância que pode ser utilizada para calcular a similaridade entre os objetos é a distância euclidiana. Já a similaridade entre os grupos pode ser calculada através de métodos como o *complete linkage* ou o UPGMA.

Figura 7 – Filogenia gerada para representar visualmente a árvore filogenética no formato *Newick* descrita como $((A:7.0,B:2.0):3.0,((C:3.0,D:3.0):2.0,E:10):8.0);$, na qual se considera o comprimento dos ramos

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

2.10.1 *Complete linkage*

No método *complete linkage*, a distância entre dois grupos é dada pela distância entre os seus objetos mais distantes, ou seja, é calculada a distância máxima entre os grupos, conforme pode se observar na equação 1:

$$d_{G_i, G_j} = \max_{g \in G_i, g' \in G_j} |g - g'| \quad (1)$$

onde d_{G_i, G_j} é a distância entre os grupos G_i e G_j e $|g - g'|$ é a distância entre os objetos (no caso deste trabalho, genomas) g , pertencente ao grupo G_i , e g' , pertencente ao grupo G_j (HAN; KAMBER, 2006).

2.10.2 UPGMA

O UPGMA (sigla para o termo, em inglês, *Unweighted Pair Group Method using Arithmetic averages*) é um método para cálculo de distâncias que pode ser utilizado para geração de agrupamentos hierárquicos. Para reconstrução de filogenias, esse método considera que as substituições de genes ocorridas nos genomas ao longo da evolução foram constantes. Desse modo, erros podem ocorrer na topologia se a taxa de substituição dos genes não for constante ou se o número de genes considerado para a geração da filogenia for pequeno (NEI; KUMAR, 2000).

Nesse método, diferentemente do método *complete linkage*, a similaridade entre dois grupos é dada pela distância média entre todos os pares de genomas de cada grupo, conforme a equação 2:

$$d_{G_i, G_j} = \frac{1}{n_i n_j} \sum_{g \in G_i, g' \in G_j} |g - g'| \quad (2)$$

onde d_{G_i, G_j} é a distância entre os grupos G_i e G_j ; n_i e n_j são os números de genomas pertencentes, respectivamente, aos grupos G_i e G_j ; e $|g - g'|$ é a distância entre os objetos (no caso, genomas) g , pertencente ao grupo G_i , e g' , pertencente ao grupo G_j .

2.11 Ferramentas e conceitos computacionais

2.11.1 BLAST

BLAST (sigla para o termo, em inglês, *Basic Local Alignment Search Tool*) é uma ferramenta de alinhamento local que busca por regiões de similaridade entre sequências de genomas. Para tanto, compara-se as sequências de nucleotídeos ou de aminoácidos com sequências presentes em um banco de dados e calcula-se a significância estatística das correspondências encontradas. A ferramenta pode ser utilizada tanto para inferência funcional quanto para inferência de relações evolucionárias entre as sequências, assim como para auxiliar na identificação de membros de famílias de genes.

O BLAST realiza alinhamentos locais de modo que é computada uma matriz com pontuações para todos os possíveis pares de segmentos de sequências, tal que identidades e substituições conservadas nas sequências possuem pontuações positivas, enquanto que substituições improváveis recebem pontuações negativas. Para tanto, a ferramenta compara todas as sequências presentes no banco de dados escolhido com a sequência fornecida

como entrada e visa a encontrar todas as sequências que, juntamente com a sequência de entrada, formam um par de segmentos máximo acima de um dado limiar. Um par de segmentos máximo (do inglês, *maximal segment pair* - MSP) é definido como o par com pontuação mais elevada entre segmentos de comprimento idêntico. A pontuação do MSP é calculada heurísticamente e fornece uma medida da similaridade local para qualquer par de sequências. Desse modo, um par de segmentos é considerado máximo se sua pontuação não puder melhorar ao se estender ou encurtar esses segmentos (ALTSCHUL et al., 1990).

O BLAST pode ser configurado para gerar arquivos tabulados no formato *m8*, que contém 12 campos com informações sobre os alinhamentos encontrados pela ferramenta. Esses campos são:

- *query name*: nome ou o identificador da sequência fornecida como entrada para a ferramenta;
- *subject name*: nome ou o identificador da sequência da base de dados que foi alinhada com a sequência de entrada;
- *percent identities*: porcentagem da ocorrência, ao longo do comprimento do alinhamento, dos mesmos nucleotídeos ou aminoácidos na mesma posição em ambas as sequências alinhadas;
- *alignment length*: comprimento do alinhamento;
- *number of mismatched positions*: número de posições em que nucleotídeos ou aminoácidos diferentes foram alinhados;
- *number of gap positions*: trata-se do número de espaços em branco (lacunas) inseridos no alinhamento;
- *query sequence start*: posição inicial do alinhamento na sequência de entrada, ou seja, onde o alinhamento começou;
- *query sequence end*: posição final do alinhamento na sequência de entrada (onde o alinhamento terminou);
- *subject sequence start*: posição inicial do alinhamento na sequência encontrada no banco de dados;
- *subject sequence end*: posição final do alinhamento na sequência encontrada no banco de dados;

- *e-value*: número esperado de alinhamentos que tenham sido obtidos ao acaso no banco de dados. Quanto menor o *e-value*, mais significativo é o *bit-score* obtido e menores são as chances de que o alinhamento encontrado tenha ocorrido pelo acaso;
- *bit-score*: nota normalizada derivada do *score*. O *score* consiste na nota dada ao alinhamento com base nos pareamentos de nucleotídeos/ aminoácidos idênticos e diferentes entre as sequências alinhadas.

Algumas variações da ferramenta foram criadas ao longo dos anos. Uma delas é o BLASTP, que realiza a comparação de uma sequência de aminoácidos fornecida como entrada com sequências presentes em um banco de dados de proteínas. Há também o TBLASTN, em que é possível realizar o alinhamento de sequências de aminoácidos com sequências de nucleotídeos presentes na base de dados consultada. Para tanto, essa versão do BLAST realiza a tradução das sequências de nucleotídeos em hipotéticas sequências de aminoácidos e alinha essas sequências hipotéticas com a sequência de entrada ([GERTZ et al., 2006](#)).

2.11.2 GGDC

O *web service* GGDC (sigla para o termo, em inglês, *Genome-to-Genome Distance Calculator*) permite a utilização do GBDP (sigla para o termo, em inglês, *Genome Blast Distance Phylogeny*) ([MEIER-KOLTHOFF et al., 2013](#)), que se trata de uma abordagem que pode ser utilizada para inferir árvores filogenéticas a partir de um dado conjunto de sequências genômicas completas ou incompletas. Essa abordagem permite estimar, computacionalmente, e obter os mesmos resultados da hibridização DNA-DNA (do inglês, *DNA-DNA hybridization* - DDH), que é uma técnica realizada em laboratório para identificar a distância entre dois genomas, tal que eles são considerados como pertencentes a uma mesma espécie se a similaridade DDH entre eles for superior a 70% ([WAYNE et al., 1987](#)).

Para tanto, dois genomas são inicialmente alinhados localmente por meio de ferramentas de alinhamento já existentes e as correspondências resultantes desse alinhamento entre os dois genomas formam um conjunto do que foi chamado de pares de segmentos de alta pontuação (do inglês, *high-scoring segment pairs* - HSPs) que, por sua vez, representam alinhamentos locais considerados estatisticamente relevantes se o seu *e-value*

correspondente for suficientemente baixo. Em seguida, informações contidas nesses pares de segmentos de alta pontuação, como o número total de pares de base idênticos entre os dois genomas alinhados, são utilizadas para se obter um valor de distância entre os dois genomas. O GGDC utiliza três fórmulas diferentes para calcular essa distância intergenômica. Na primeira fórmula, a distância é dada pelo comprimento total dos HSPs dividido pela soma dos comprimentos de ambos os genomas; na segunda, a distância é obtida pela soma de todos os pares de base idênticos existentes em todos os HSPs dividida pelo comprimento total dos HSPs; por fim, na terceira e última fórmula, a distância é dada pela soma de todos os pares de base idênticos existentes em todos os HSPs dividida pela soma dos comprimentos de ambos os genomas. Segundo os autores, a terceira fórmula é a que conserva mais informação por ser uma combinação das outras fórmulas e é a que apresenta o melhor desempenho do ponto de vista filogenético. Entretanto, os autores recomendam a utilização da segunda fórmula por ela não depender do comprimento dos genomas e por ser a única que pode ser utilizada para genomas incompletos.

Com base nessas distâncias calculadas para os genomas, pode-se criar uma matriz de distância para inferir e gerar árvores filogenéticas.

Além disso, a ferramenta também permite estimar o valor DDH a partir das distâncias calculadas. Para tanto, utiliza um modelo estatístico chamado de modelo linear generalizado (NELDER; BAKER, 1972), tal que os valores DDH são considerados variáveis resposta e os correspondentes valores das distâncias entre os genomas são as variáveis preditoras.

O modelo linear generalizado também foi modificado de modo que a variável resposta foi convertida para uma variável resposta binária sobre a proporção de DDH, tal que, para uma dada distância entre genomas, ele retorna a probabilidade de que, de fato, ela corresponda ou não a um valor DDH maior do que 70%.

2.11.3 MUSCLE

MUSCLE (sigla para o termo, em inglês, *MUltiple Sequence Comparison by Log-Expectation*) é uma ferramenta desenvolvida para realizar o alinhamento múltiplo de sequências de aminoácidos e nucleotídeos. Esse problema pode ser formulado computacionalmente de modo que se define um modelo da evolução da sequência, calcula-se as

probabilidades das mudanças das sequências e, por meio de heurísticas, busca-se pelo grafo direcionado mais provável, em que as arestas representam as mudanças e os nós terminais são as sequências observadas.

O algoritmo da ferramenta é composto por três etapas principais. Na primeira, tem-se como objetivo produzir o alinhamento múltiplo, priorizando a rapidez em detrimento da acurácia. Primeiramente, uma matriz de distância é obtida a partir do cálculo das distâncias *kmer* para cada par de sequências, em que um *kmer* consiste em uma subsequência contínua de tamanho k e tal que sequências relacionadas tendem a possuir mais *kmers* em comum do que o que é esperado pelo acaso. Com isso, considera-se a fração de *kmers* que as sequências têm em comum para realizar o cálculo da distância. Em seguida, a matriz é agrupada com o método UPGMA, resultando em uma árvore binária. O alinhamento é então realizado de acordo com a ordem dos ramos presentes na árvore binária obtida. Para cada folha, um perfil, que consiste em um alinhamento múltiplo tratado como uma sequência e considerando cada coluna como um símbolo alinhável, é construído para uma sequência de entrada. Os nós da árvore são visitados em pré-ordem (filhos antes de seus pais) e, para cada nó interno, um alinhamento de pares é construído com os dois perfis dos filhos, produzindo um novo perfil que é atribuído àquele nó.

Na segunda etapa, a árvore é re-estimada utilizando a distância Kimura (KIMURA, 1983), visto que a distância *kmer* pode gerar uma árvore sub-ótima. A distância Kimura, com base na identidade entre os pares de sequências, calcula uma distância estimada ao aplicar uma correção para as múltiplas substituições ocorridas em um único local, o que, apesar de fornecer maior acurácia, necessita que um alinhamento tenha sido realizado previamente. Primeiramente, a distância Kimura é calculada para cada par de sequências de entrada a partir do alinhamento múltiplo realizado na etapa anterior, resultando em uma matriz de distância. Essa matriz também é agrupada com o UPGMA e produz uma árvore binária. Um alinhamento progressivo é realizado e produz-se o alinhamento múltiplo. Esse processo é otimizado ao calcular os alinhamentos apenas das sub-árvores cuja ordem dos ramos se alterou com relação à árvore binária obtida na primeira etapa.

A terceira e última etapa consiste no refinamento do alinhamento múltiplo. Primeiramente, uma aresta é escolhida na segunda árvore binária. Essa árvore é dividida em duas sub-árvores ao excluir a aresta escolhida. O perfil do alinhamento múltiplo de cada sub-árvore é obtido. Um novo alinhamento múltiplo é realizado ao realinhar os dois perfis. Se a soma das pontuações dos alinhamentos entre pares melhorar, o novo alinhamento é

guardado ou descartado, caso contrário. Esses passos são repetidos até a convergência ou até que o limite definido pelo usuário seja atingido (EDGAR, 2004).

2.11.4 PhyML

PhyML é uma ferramenta desenvolvida para gerar inferências filogenéticas a partir de alinhamentos de sequências de aminoácidos ou de nucleotídeos baseando-se no princípio da máxima verossimilhança. A versão mais recente da ferramenta, que foi utilizada no presente trabalho, utiliza o algoritmo chamado de NNI (sigla para o termo, em inglês, *Nearest Neighbor Interchanges*) como método padrão para se obter a árvore filogenética com maior verossimilhança.

Nesse algoritmo, primeiramente, uma matriz de distância é calculada a partir da comparação entre todos os pares de sequências de entrada e uma árvore inicial é construída a partir dessa matriz. As verossimilhanças de cada subárvore e da árvore como um todo são então calculadas e os valores dos parâmetros do modelo de substituição da árvore são então ajustados de modo a aumentar a verossimilhança da filogenia inicial. Em seguida, a árvore atual é refinada iterativamente até sua convergência, tal que esse refinamento consiste em avaliar possíveis modificações em cada ramo da árvore, aplicar algumas dessas modificações na árvore atual e verificar se essas mudanças ocasionaram um aumento da verossimilhança da árvore. O refinamento da árvore é finalizado quando não há mais nenhuma mudança na árvore ou quando os comprimentos de seus ramos se estabilizarem. Com isso, a árvore filogenética resultante é retornada (GUINDON; GASCUEL, 2003).

2.11.5 PHYLIP

PHYLIP (sigla para o termo, em inglês, *PHYLogeny Inference Package*) é uma ferramenta para inferência de filogenias (FELSENSTEIN, 1989). Diversos métodos estão disponíveis para realizar essa inferência, dentre os quais estão os que utilizam matrizes de distância para realizar esta tarefa. Em um dos experimentos apresentados neste trabalho, utilizou-se o método UPGMA, presente no pacote *neighbor*. Após a utilização do UPGMA, a ferramenta produz como saída uma filogenia no formato *Newick*.

2.11.6 R

O R é um *software* desenvolvido para a realização de cálculos estatísticos e geração de gráficos (R Core Team, 2016). Ele apresenta diversos pacotes, que contém um conjunto de funções que podem ser utilizadas para fins específicos. Neste trabalho, dois pacotes foram utilizados: o *stats* e o *ape*.

O pacote *stats* (R Core Team, 2016) contém a função *hclust*, que permite a realização de agrupamentos hierárquicos. O método padrão e que foi utilizado no presente trabalho para realização dos agrupamentos é o *complete linkage*. Como entrada para essa função, deve-se fornecer uma matriz de distância dos objetos que serão agrupados.

Já o pacote *ape* (PARADIS; CLAUDE; STRIMMER, 2004) foi utilizado por conta da função *read.tree*. Essa função permite a leitura de arquivos contendo árvores no formato *Newick* para posterior geração de dendrogramas para representação visual das árvores filogenéticas. Como entrada, deve-se fornecer o caminho para o arquivo ou o texto contendo a árvore no formato *Newick*.

3 Trabalhos correlatos

Este capítulo tem o intuito de apresentar trabalhos relacionados com o projeto exposto nesta dissertação. Para tanto, uma revisão bibliográfica sistemática foi realizada com o objetivo de identificar o estado da arte na comparação de genomas completos de bactérias por meio de técnicas e métodos computacionais. Além dessa revisão, também foi realizada uma revisão bibliográfica não sistemática em busca de trabalhos sobre métodos para se realizar a inferência filogenética. As revisões sistemática e não sistemática são apresentadas, respectivamente, nas seções 3.1 e 3.2.

3.1 Revisão sistemática

Uma revisão sistemática da literatura é uma metodologia de busca bibliográfica rigorosa e confiável que consiste na documentação dos passos seguidos pelos revisores, permitindo que a mesma seja replicada (KITCHENHAM, 2004; BIOLCHINI *et al.*, 2005).

Por esse motivo, uma revisão sistemática foi realizada visando à contextualização do projeto proposto com o que têm sido desenvolvido na área de bioinformática. Ela foi conduzida de acordo com as três fases definidas por Kitchenham (2004) e Biolchini *et al.* (2005): planejamento, condução e extração de dados. Na primeira fase, foi definido um protocolo, no qual foi especificado como a revisão seria conduzida. Na segunda, foi executada a busca por trabalhos considerados relevantes de acordo com os critérios de inclusão e exclusão definidos no protocolo. Na terceira e última fase, foi realizada a extração e documentação dos dados dos trabalhos considerados relevantes, que, por meio de sua análise, permitiram obter uma visão geral do estado da arte na comparação de genomas completos de bactérias.

A revisão sistemática teve como intuito identificar e analisar métodos e técnicas computacionais já existentes para a comparação de genomas completos de bactérias. Dessa forma, ao final da revisão sistemática, obteve-se uma visão do estado da arte na comparação de genomas completos de bactérias, bem como de oportunidades para desenvolvimento de métodos e técnicas para realizá-la.

Os resultados obtidos e as análises realizadas ao final da revisão são descritos a seguir. O protocolo elaborado na fase de planejamento e as especificações de como a condução foi realizada estão disponíveis *online* ^{1 2}.

As buscas pelos artigos foram realizadas no dia 31 de julho de 2015 nas bases de dados da IEEE³ e Science Direct⁴ e nos dias 4 e 6 de setembro de 2015 nas bases da PLOS ONE⁵ e ACM⁶, respectivamente. Por conta das particularidades das ferramentas de busca de cada base de dados, foram construídas *strings* de busca específicas para cada uma delas. Além disso, para as bases PLOS ONE e Science Direct, que não são bases específicas da área de computação, termos computacionais foram acrescentadas nas *strings* de busca.

No total, 208 trabalhos foram recuperados pelas buscas e seus *abstracts* foram lidos para aplicação dos critérios de inclusão e exclusão descritos no protocolo elaborado na fase de planejamento. Após a leitura e aplicação dos critérios, 23 trabalhos foram incluídos na revisão sistemática. Por fim, esses 23 trabalhos foram lidos integralmente e os critérios de inclusão e exclusão foram aplicados novamente. Com isso, 15 trabalhos foram analisados na revisão sistemática, sendo que dois correspondem às revisões e comparações de ferramentas que realizam a comparação de genomas e os 13 demais apresentam métodos e técnicas para comparação de genomas completos de bactérias. A figura 8 apresenta a distribuição dos artigos inicialmente encontrados nas quatro bases de dados, bem como os selecionados ao final da fase de condução da revisão sistemática.

A figura 9, por sua vez, apresenta o número de artigos publicados de acordo com o ano em que foram publicados. Por meio dela, é possível notar o pico formado no ano de 2009, em que um total de seis artigos foram publicados, o que pode indicar que pesquisadores estavam focados no desenvolvimento de métodos e técnicas para a comparação de genomas completos de bactérias.

Já a figura 10 apresenta o número e a porcentagem de artigos de acordo com as categorias definidas para distinguir os métodos e técnicas identificados. Por meio dela, é possível observar que a distribuição pelas categorias foi, de certa forma, equilibrada, sendo que as árvores foram as mais utilizadas (34% dos trabalhos) para realizar a comparação de genomas ao longo dos anos.

¹ <http://www.each.usp.br/digiampietri/Xmeeting2015/PereiraEtAlProtocol.pdf>

² <http://www.each.usp.br/digiampietri/Xmeeting2015/PereiraEtAlConduct.pdf>

³ <http://ieeexplore.ieee.org/Xplore>

⁴ <http://sciencedirect.com>

⁵ <http://plosone.org>

⁶ <http://portal.acm.org>

Figura 8 – Distribuição dos artigos encontrados e selecionados em cada base de dados

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

Figura 9 – Distribuição dos artigos encontrados de acordo com o ano de publicação

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

Como mencionado anteriormente, somente dois artigos encontrados tratavam-se de revisões sobre métodos e técnicas para comparação de genomas completos de bactérias.

Altenhoff e Dessimoz (2009) compararam 11 projetos que objetivavam a identificação de genes ortólogos em diferentes organismos. Entretanto, alguns deles não puderam ser avaliados com relação à sua aplicação para bactérias por possuírem poucos dados apropriados. No experimento que visava a avaliar a especificidade das ferramentas ao identificar os genes ortólogos de bactérias, árvores filogenéticas foram reconstruídas com base nos genes ortólogos identificados e essas árvores resultantes foram comparadas com uma filogenia de referência. Os pesquisadores observaram que as ferramentas COG (TATUSOV; KOONIN; LIPMAN, 1997), EggNOG (JENSEN et al., 2008) e OrthoMCL (LI;

Figura 10 – Distribuição dos artigos encontrados de acordo com a categoria para a qual foram classificados

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

STOECKERT; ROOS, 2003) apresentaram altas taxas de falsos positivos, o que ocasionou em um número reduzido de árvores reconstruídas corretamente. Como o conjunto de sequências utilizados por cada projeto é diferente e como o conjunto de intersecção dessas sequências não é suficientemente numeroso, não foi possível comparar de maneira precisa as ferramentas que obtiveram os melhores resultados para os genomas de bactérias, que foram as ferramentas OMA (DESSIMOZ et al., 2005) e RoundUp (DELUCA et al., 2006).

Laing et al. (2011), por sua vez, apresentam uma revisão sobre a comparação de genomas completos de bactérias, em que são mencionadas ferramentas existentes, bem como seus desempenhos no alinhamento de sequências genômicas, identificação de duplicações e rearranjos em genomas. O artigo cita que árvores de sufixo foram utilizadas pela ferramenta MUMmer (KURTZ et al., 2004) para realizar o alinhamento de genomas completos de maneira eficiente e que a ferramenta MISHIMA (KRYUKOV; SAITOU, 2010), que utiliza a heurística de divisão e conquista para realizar o alinhamento múltiplo de genomas completos, é mais rápida do que métodos de comparação entre pares de sequências, mas que apresenta bons resultados apenas ao comparar genomas completos evolutivamente próximos.

Os 13 artigos que não se tratam de revisões foram classificados em categorias de acordo com as técnicas e métodos para a comparação de genomas de bactérias apresentados: a categoria “I - Grafos” compreende os métodos e técnicas que usaram grafos como estrutura

de dados para representar e comparar os genomas; “II - Árvores” compreende os métodos e técnicas que usaram árvores como estrutura de dados para representar e comparar os genomas; a categoria “III - Aprendizado de máquina” apresenta os métodos e técnicas que usaram abordagens de aprendizado supervisionado ou não-supervisionado para comparar, classificar ou identificar características em genomas; e, por fim, a categoria “IV - Outros métodos” contempla os métodos e técnicas que não se encaixaram nas categorias anteriores. A tabela 1 apresenta um resumo dos dados extraídos dos 13 artigos, bem como as categorias para as quais cada um foi classificado.

As próximas subseções detalham os artigos em suas respectivas categorias.

3.1.1 I - Grafos

Apenas os trabalhos de [Lam et al. \(2007\)](#) e [Hachiya et al. \(2009\)](#) utilizaram grafos para a comparação dos genomas.

[Lam et al. \(2007\)](#) desenvolveram algoritmos para representar padrões estruturais de genomas e compará-los por meio de um grafo multinível utilizado para encontrar similaridades entre os genomas. Os padrões estruturais foram representados no primeiro nível do grafo, tal que os nós eram os genes e as arestas eram as sequências de nucleotídeos não codificantes que representavam as relações de conexão entre dois genes. Nos níveis mais elevados, os nós eram os segmentos de genomas e as arestas eram as relações de adjacência dos segmentos de genes conectados e as múltiplas relações de genes correspondentes. A ferramenta desenvolvida foi testada com genomas completos de bactérias e as comparações realizadas indicaram quais genomas eram mais próximos evolutivamente.

Já [Hachiya et al. \(2009\)](#) desenvolveram um algoritmo para mapear e identificar segmentos ortólogos. Para tanto, utilizaram um grafo direcionado acíclico para realizar a comparação entre os genomas, tal que os nós do grafo eram as pequenas regiões bem conservadas entre os genomas, denominadas *anchors*, e as arestas eram *anchors* colineares. Também foi utilizada a distribuição geométrica definida pelo modelo de cadeia de Markov para distinguir *anchors* ortólogos de não ortólogos e também para fornecer uma pontuação para as arestas do grafo. Com base nessa pontuação, utilizou-se a técnica de programação dinâmica para extrair caminhos sub-ótimos que não se intersectassem e, com isso, um conjunto de cadeias foi identificado. Por fim, uma sequência de cadeias colineares foi

Tabela 1 – Resumo das informações obtidas nos artigos incluídos na revisão sistemática

Artigo	Método/ técnica	Objetivo	Comparação	Validação	Categoria
Landau, Parida e Weimann (2005)	Árvore PQ	Detecção de agrupamentos de genes	Múltiplos genomas	-	II
Angibaud et al. (2006)	Abordagem pseudo-booleana	Encontrar o número mínimo de <i>breakpoints</i>	Em pares	Outros estudos	IV
Goto, Kurokawa e Yasunaga (2007)	Árvore de sufixo	Identificação de sequências conservadas em todos os genomas	Múltiplos genomas	Outros estudos	II
Lam et al. (2007)	Grafo multinível	Comparação de estruturas de genomas	Em pares	Outros estudos	I
Hachiya et al. (2009)	Grafo direcionado acíclico	Mapeamento de ortologias	Múltiplos genomas	F-score, sensibilidade, especificidade e anotação de ortologias	I
Domazet-Lošo e Haubold (2009)	Árvore de sufixo e <i>average shortest unique substring</i>	Cálculo de distâncias para reconstrução de filogenia	Múltiplos genomas	Outros estudos	II
Bolshoy e Volkovich (2009)	Aprendizado não-supervisionado e matriz esparsa	Reconstrução filogenética	Múltiplos genomas	Taxonomia do NCBI	III
Das, Vishnoi e Bhattacharya (2009)	Árvore de sufixo	Identificação de variações genômicas	Em pares	-	II
Kozobay-Avraham et al. (2009)	<i>K-means</i> e distância quadrada euclidiana	Comparação e classificação de procariontes	Múltiplos genomas	Genomas do NCBI, o índice Krzanowski e Lai e o índice Sugar e James	III
Darling, Mau e Perna (2010)	Modelo oculto de Markov (HMM)	Alinhamento de genoma	Múltiplos genomas	Valor de predição positiva, curadoria manual e modelos de simulação evolucionária	III
Jahn (2010)	Abordagem <i>common intervals</i>	Detecção de agrupamento de genes	Múltiplos genomas	Outro estudo	IV
Comin e Verzotto (2012)	Árvore de sufixo e abordagem <i>UNDERlying-paired Irredundant Common subwords (unic subwords)</i>	Reconstrução filogenética	Em pares	Robinson-Foulds e filogenia baseada no gene 16S rRNA	II
Kubicova e Provaznik (2014)	Análise de espectrograma	Reconstrução filogenética	Em pares	Coeficiente de correlação de Matthew e taxonomia do NCBI	IV

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

combinada para se obter os segmentos ortólogos. O algoritmo foi testado utilizando genomas completos de bactérias e obteve-se um F-score superior a 90% na comparação entre múltiplos genomas de bactérias. Além disso, o F-score médio do algoritmo foi maior na comparação de múltiplos genomas do que apenas entre dois genomas.

3.1.2 II - Árvores

Dentre todos os métodos encontrados nos artigos selecionados, o uso de árvores como estrutura de dados para realizar a comparação de genomas foi o mais frequente, com cinco dos 13 trabalhos utilizando-as e, desses cinco, quatro utilizaram árvores de sufixo.

Goto, Kurokawa e Yasunaga (2007) utilizaram árvore de sufixo para identificar sequências conservadas ao longo de todos os genomas, tanto em sequências de DNA quanto de aminoácidos. Após a criação da árvore, os nós internos foram examinados e as sequências conservadas que satisfizessem o limiar estabelecido pelo usuário eram reportadas. O software criado foi testado com 143 genomas, dos quais 125 eram de bactérias, e a hipótese dos autores era de que as sequências conservadas encontradas estavam presentes no último ancestral em comum de todas as formas de vida.

Das, Vishnoi e Bhattacharya (2009) desenvolveram uma ferramenta para identificar inserções, deleções e inversões entre dois genomas. Depois da identificação e extração das sequências de regiões bem conservadas (*anchors*) do genoma de referência, o algoritmo baseado em árvore de sufixo foi utilizado para identificar as inversões dos genomas, assim como translocações, recombinações e largas inversões. A ferramenta foi testada com bactérias e notaram que o número de variantes dependia da espécie estudada.

Domazet-Lošo e Haubold (2009) também utilizaram árvore de sufixo e o conceito de *average shortest unique substring* para calcular a distância entre sequências de DNA homólogas. O método desenvolvido obteve melhores resultados com genomas de *Drosophila* e do HIV (sigla para o termo, em inglês, *Human Immunodeficiency Virus*) do que com genomas de bactérias.

Comin e Verzotto (2012) propuseram uma medida de distância para comparação de genomas para reconstrução filogenética de genomas completos. Para tanto, o método utilizava a árvore de sufixo para identificar subpalavras únicas e calcular a distância entre os genomas. O método foi testado com genomas de procariotos e as filogenias geradas

foram comparadas com a que se baseava no gene *16S rRNA*. De acordo com os testes, o método distinguiu bem os organismos do reino *Archea* dos pertencentes ao reino *Bacteria*.

Landau, Parida e Weimann (2005), por sua vez, não usaram árvore de sufixo, mas a árvore PQ (BOOTH; LUEKER, 1976) (que pode representar possíveis permutações de um conjunto de acordo com restrições estabelecidas e cujos nós internos são denominados de P ou Q , tal que o nó P tem pelo menos dois filhos e o nó Q tem pelo menos três) como estrutura de dados para identificar agrupamentos de genes. Essa estrutura foi utilizada para descrever as relações entre agrupamentos de genes e permitir que grupos irrelevantes fossem filtrados por meio da identificação de uma árvore PQ de consenso mínimo, que é a representação de uma notação máxima de k ocorrências de um π padrão. Dado k permutações em um alfabeto (que no caso, são os genes) representando k ocorrências de um padrão e sejam um padrão e uma *string* (que representa um genoma) sequências de caracteres do alfabeto, a notação máxima é obtida ao usar um hífen entre dois grupos de um ou mais genes para denotar que esses grupos aparecem como vizinhos imediatos em todas as k permutações ou uma vírgula, caso contrário. Já um π padrão é um padrão que aparece ao menos k vezes na *string* para um dado k . Um dos experimentos realizados utilizou os genomas das bactérias *Escherichia coli K-12* e *Bacillus subtilis*. Os autores apenas citam que 450 árvores PQ de consenso mínimo de 15.000 agrupamentos de genes foram encontradas e que o gene COG0718, cuja função não é determinada, foi mapeado entre dois genes cuja função está relacionada com a reparação de DNA e, com isso, afirmam que essa informação pode indicar que esse gene também pode estar relacionado com essa função.

3.1.3 III - Aprendizado de máquina

Dos 13 artigos encontrados, três utilizam técnicas de aprendizado supervisionado e não-supervisionado para realizar a comparação de genomas completos.

Bolshoy e Volkovich (2009) propuseram um método de agrupamento de organismos unicelulares baseado no comprimento de todos os genes ortólogos para posterior representação dos grupos encontrados em uma árvore filogenética. Para isso, utilizou-se uma matriz esparsa contendo a média do comprimento das proteínas ortólogas dos genomas e o método *information bottleneck* para medir a similaridade entre dois genomas e realizar

o agrupamento aglomerativo e não supervisionado dos mesmos, de modo que fossem encontradas partições que maximizassem a informação mútua entre os genomas. O teste do método demonstrou que ele foi capaz de separar os reinos *Archea* e *Bacteria*.

Já [Kozobay-Avraham et al. \(2009\)](#) utilizaram distribuições da curvatura do DNA para comparar e classificar procariotos. Para tanto, fizeram uma análise de agrupamentos utilizando o *k-means* e o PAM (sigla para o termo, em inglês, *Partition Around Medoids*), assim como seis diferentes medidas de distância para descobrir a melhor combinação de técnica de agrupamento e medida de distância, isto é, a mais razoável do ponto de vista biológico. Análises realizadas demonstraram que a combinação do *k-means* com a distância euclidiana entre os genomas representou melhor os genomas do ponto de vista biológico. Os pesquisadores utilizaram 205 genomas completos de procariotos, dos quais 182 eram de bactérias, e os genomas foram separados em três grupos. Eles observaram uma correlação entre a taxonomia utilizada como referência e os agrupamentos encontrados com o *k-means*.

[Darling, Mau e Perna \(2010\)](#) desenvolveram um método de alinhamento para alinhar regiões conservadas de genomas. Para tanto, o modelo oculto de Markov (do inglês, *Hidden Markov Model* - HMM) foi utilizado para prever homologias e para identificar as sequências não relacionadas que foram alinhadas incorretamente e descartá-las do alinhamento final. Os autores apresentaram algumas aplicações do método, como a identificação da variabilidade de conteúdo gênico em espécies de bactérias.

3.1.4 IV - Outros métodos

Três trabalhos não se encaixaram nas categorias apresentadas anteriormente por utilizarem métodos diferentes para comparar os genomas.

[Angibaud et al. \(2006\)](#) utilizaram uma abordagem pseudo-booleana para encontrar a similaridade e a distância genômica entre dois genomas completos. Para tanto, procurou-se identificar o número máximo de *common intervals* entre dois genomas utilizando o modelo *maximum matching*, que consiste em mapear todos os genes possíveis. No teste realizado com a abordagem, apenas 39 das 66 combinações possíveis de comparações entre dois genomas de bactérias foram obtidas devido ao seu elevado tempo de processamento.

[Jahn \(2010\)](#) implementou um algoritmo para identificar agrupamentos de genes, por meio da análise de segmentos bem conservados nos genomas. O algoritmo utilizou

um arranjo para encontrar *common intervals* entre os genomas. Ele foi comparado com outra abordagem para identificação de agrupamentos de genomas e ambos apresentaram bons resultados na comparação de genomas de bactérias e tiveram como fator limitante o número de genomas fornecidos como entrada.

Por fim, [Kubicova e Provaznik \(2014\)](#) apresentaram um método de comparação de genomas completos baseado na comparação de espectogramas, que possibilitaram uma visão da periodicidade local em sequências de nucleotídeos e seus padrões relacionados à estrutura ou função. A distância entre os genomas foi medida pela similaridade entre seus espectogramas, que foi calculada pela distância euclidiana. A validação da ferramenta utilizando genomas de bactérias demonstrou que classificações produzidas para o táxon ordem apresentou melhores resultados do que considerando o filo dos organismos.

3.1.5 Considerações finais sobre a revisão sistemática

Primeiramente, é importante salientar que uma possível limitação dessa revisão sistemática e que pode ter se refletido no baixo número de artigos incluídos na revisão é que alguns artigos foram excluídos durante a fase de condução, pois, apesar de realizarem a comparação de genomas completos, não apresentavam aplicações dos métodos e técnicas em genomas de bactérias. Por não terem explicitado se a ferramenta poderia ser utilizada na comparação de bactérias e por não possibilitarem analisar seus resultados no que diz respeito a esses genomas, os artigos não foram incluídos nessa revisão sistemática.

Analisando os artigos obtidos durante a revisão sistemática, notou-se que cerca da metade deles apresentou métodos ou técnicas para realizar a comparação apenas entre dois genomas e não entre múltiplos genomas. Essa pode ser considerada uma limitação pelo fato de que a comparação entre apenas dois genomas pode prejudicar a detecção de sequências conservadas com acurácia e, apesar da maior complexidade na comparação entre múltiplos genomas, ela apresenta resultados mais precisos sobre similaridades e divergências entre genomas de bactérias ([FRAZER, 2003](#)).

Também foi possível observar que os trabalhos procuraram distinguir genomas de organismos de táxons como gênero, ordem e espécie, mas não de organismos muito próximos evolutivamente, como subespécies ou diferentes cepas e patovares de uma espécie. Apesar da dificuldade para diferenciá-los, o estudo desses organismos é interessante para o

entendimento e comparação de suas particularidades, que podem auxiliar em pesquisas sobre doenças causadas por eles, por exemplo.

Além disso, observou-se que alguns estudos validaram seus métodos ao comparar seus resultados com a filogenia de referência baseada no gene *16S rRNA* que, apesar de ser amplamente utilizado, tem um baixo poder filogenético para distinguir espécies muito próximas evolutivamente (JANDA; ABBOTT, 2007). Por esse motivo, é importante verificar outras filogenias além da baseada no gene *16S rRNA* para os casos em que se deseja realizar a reconstrução filogenética para organismos evolutivamente próximos, tais como os mencionados anteriormente.

Por fim, também foi possível notar que muitos dos trabalhos fizeram a comparação dos genomas com base nos genes homólogos e ortólogos ou objetivaram a identificação desses genes. Isso pode ter ocorrido pelo fato de que, de acordo com sua definição, dois genes homólogos são ortólogos se os mesmos divergem após um evento de especiação e, dessa forma, uma árvore filogenética gerada a partir de uma família de genes ortólogos (um conjunto de genes em que qualquer par é ortólogo) tem, provavelmente, a mesma topologia da correspondente árvore filogenética de espécies (ALTENHOFF; DESSIMOZ, 2009). Por esse motivo, no presente trabalho também procurou-se identificar esses genes para utilizá-los na reconstrução filogenética.

3.2 *Revisão bibliográfica não sistemática*

Nesta seção, serão apresentados trabalhos encontrados a partir de uma revisão bibliográfica não sistemática que objetivou a procura por trabalhos sobre a reconstrução filogenética. Os trabalhos encontrados foram divididos nas próximas subseções de acordo com os métodos para reconstrução filogenética baseados em sequências e baseados em características de genomas completos, descritos na seção 2.8.

3.2.1 Filogenia baseada em sequências

Daubin, Gouy e Perriere (2002) adotaram o método de superárvore para realizar a reconstrução filogenética de 41 procariotos e quatro eucariotos. Para tanto, foram consideradas as famílias de genes homólogos presentes nos genomas, de modo que os

autores mantiveram somente as famílias que continham apenas um gene em cada espécie. Em seguida, cada família de genes foi alinhada e uma árvore de máxima verossimilhança (ML) foi gerada para cada família, assim como também foi gerada uma árvore utilizando o algoritmo BIONJ (GASCUEL, 1997), que é uma versão aperfeiçoada do algoritmo *neighbor-joining*. Cada uma dessas árvores foi então computada em uma matriz binária e, por fim, as matrizes foram concatenadas em uma supermatriz e a superárvore foi gerada a partir dela.

Tanto a árvore reconstruída com o método de máxima verossimilhança quanto com o algoritmo BIONJ se dividiram em arqueobactérias, bactérias e eucariotos, sendo que as relações entre as arqueobactérias ficaram menos claras na árvore do BIONJ do que na árvore de máxima verossimilhança. Com relação às bactérias, as relações entre os ramos mais internos não foram muito bem resolvidas em ambas as árvores, sendo que a árvore ML apresentou uma topologia mais anormal pelo fato das ϵ -proteobactérias estarem mais distantes das demais proteobactérias e por estarem localizadas mais no início da divisão da árvore. De acordo com os autores, essa dificuldade na resolução de tais relações podem estar relacionadas com o aumento da probabilidade de ocorrência de transferências horizontais, paralogias escondidas ou devido à atração de ramificações longas, em que grupos distantes acabaram, erroneamente, sendo considerados evolutivamente próximos.

Para tentar contornar as possíveis transferências horizontais, os pesquisadores também compararam árvores geradas para cada família de genes selecionadas através da análise de coordenadas principais (do inglês, *principal coordinates analysis* - PCO) e que continham ao menos 10 espécies de bactérias. Com elas, superárvores ML e BIONJ foram geradas. Assim como as árvores anteriores, essas conseguiram separar arqueobactérias, bactérias e eucariotos. A baixa resolução que foi observada nas relações entre arqueobactérias ocorreu porque genes presentes apenas nesses organismos ou que eram compartilhados apenas por eucariotos e arqueobactérias foram removidos neste experimento por não estarem presentes em bactérias. As relações entre bactérias, por sua vez, apresentaram elevados valores de *bootstrap* e aparentaram uma melhor resolução neste experimento do que no anterior. Neste experimento, as ϵ -proteobactérias foram agrupadas com as demais proteobactérias e com valores *bootstrap* significativamente altos. Entretanto, essas árvores apresentaram diferenças substanciais no posicionamento de bactérias *Aquifex*, *Thermotoga* e *Deinococcus* com relação às filogenias geradas com base nos genes de *rRNA*, em que esses organismos são apresentados próximos à raiz da árvore. Os autores concluíram

que, apesar da análise do PCO ter reduzido o comprimento da supermatriz, os valores de *bootstrap* aumentaram na maioria dos nós de bactérias e que isso revela que o grupo de genes selecionados pela análise continha informações relevantes para a filogenia de bactérias. Além disso, afirmam que a superárvore pode ser um bom método para a inferência de filogenias pois considera a informação filogenética de centenas de genes, mas que esse método só é significativo se os genes considerados permanecerem estáveis ao longo da evolução.

Queiroz e Gatesy (2007) apresentaram uma revisão sobre as vantagens e desvantagens de se utilizar a abordagem de supermatriz para realizar a inferência filogenética. Segundo eles, as supermatrizes consideram todas as informações inseridas na matriz para gerar a filogenia, enquanto que, nas superárvores, algumas dessas informações podem ser perdidas quando elas são resumidas nas árvores. Além disso, as análises de supermatrizes podem relevar informações que apoiem relações em uma árvore, o que pode não ocorrer quando elas são vistas isoladamente, e que essas análises tendem a produzir árvores em que as relações são melhor resolvidas do que nas árvores produzidas por superárvores.

Como desvantagem, citam que a abordagem padrão de supermatrizes assume, de maneira implícita, que todas as características consideradas passaram pela mesma história evolutiva. Para contornar esse problema, outros métodos foram desenvolvidos para analisar supermatrizes, mas os autores salientam que estudos e simulações a respeito dos resultados obtidos para a filogenia com base nesses métodos são necessários para analisar se as melhorias proporcionadas geram resultados melhores dos que os obtidos pela abordagem padrão.

Já Lang, Darling e Eisen (2013) utilizaram os métodos de superárvore e supermatriz para reconstruir a filogenia de bactérias e arqueobactérias, de modo que os pesquisadores a inferiram utilizando 24 genes presentes apenas uma vez na maioria dos genomas. Após a realização dos alinhamentos desses genes, os dois métodos foram aplicados: no método de supermatriz, os alinhamentos foram concatenados para permitir uma única análise por meio da ferramenta RAxML (STAMATAKIS, 2006), pela qual se gerou uma árvore de máxima verossimilhança; já no método superárvore, os alinhamentos foram utilizados individualmente para inferir uma árvore para cada gene e, em seguida, a ferramenta BUCKy (que implementa a chamada “análise de concordância bayesiana” e que leva em conta a incerteza em uma árvore de genes e não assume que todos os genes compartilham uma mesma história evolutiva) (LARGET et al., 2010) foi utilizada para combinar essas

árvores e gerar uma única filogenia para os organismos estudados. As árvores resultantes foram então comparadas com a filogenia gerada com base no gene *16S rRNA*.

Como resultado, observou-se que as árvores geradas a partir dos métodos superárvore e supermatriz eram mais similares uma com a outra do que com a árvore gerada a partir do gene *16S rRNA*. Apesar das diferenças nos arranjos dos genomas nas três árvores, a maior parte dos agrupamentos de filos de bactérias ficaram semelhantes em todas as árvores. Por fim, os autores concluíram que a árvore de máxima verossimilhança seria a que melhor representaria a relação entre os genomas de bactérias utilizados porque permitiu gerar uma árvore completamente resolvida e que foi mais rápida de ser gerada do as das outras abordagens testadas.

É importante ressaltar que os autores removeram os genomas muito próximos evolutivamente do conjunto utilizado para realizar a reconstrução filogenética, de modo que apenas um genoma foi utilizado para representar esses organismos muito próximos. Isso porque afirmaram que esses grupos de genomas proporcionariam grandes níveis de incerteza nas análises da filogenia, visto que os genes considerados para a reconstrução filogenética seriam muito conservados, e que eles não trariam informações relevantes que contribuiriam para a inferência dos relacionamentos entre organismos muito próximos.

3.2.2 Filogenia baseada em características de genomas completos

Snel, Bork e Huynen (1999) fizeram a reconstrução filogenética com base no conteúdo gênico. Para isso, os autores consideraram os genes compartilhados pelos pares de genomas e selecionaram os pares de sequências homólogas de acordo com o limiar dado por um *e-value* igual a 0,01 (os autores citam que o *e-value* é um bom indicador para a detecção de homologias). Em seguida, foram selecionados os pares de genes de modo que o nível de identidade entre os dois genes fosse o mais alto quando comparado com o nível de identidade de cada um dos dois genes com qualquer um dos outros genes presentes no outro genoma. Após essa identificação dos genes, a distância entre cada par de genomas foi dada pela proporção de genes compartilhados por eles dividido pelo tamanho do menor dos genomas. Em seguida, utilizou-se o algoritmo *neighbour-joining* para geração da filogenia, de forma que a árvore resultante ficou similar à filogenia gerada com o *16S rRNA* com altos valores *bootstrap* de modo geral, o que, para os autores, indica que o conteúdo gênico

apresenta uma alta assinatura filogenética e que isso pode implicar que há ocorrência de poucas transferências horizontais de genes; ou que esses eventos ocorrem geralmente em espécies muito próximas ou que afetam essas espécies da mesma maneira; ou que os genes transferidos geralmente substituem genes ortólogos que já estavam presentes nos genomas. Entretanto, como o número de genomas estudados era pequeno, os autores não fizeram um estudo para analisar a probabilidade de transferência horizontal, duplicação ou perda de genes.

Já [House e Fitz-Gibbon \(2002\)](#) realizaram um estudo de conteúdo gênico com genomas de 27 espécies, dos quais 15 eram de micro-organismos. Primeiramente, matrizes foram construídas a partir da análise dos grupos de genes baseando-se na similaridade entre os pares de genomas. Para tanto, cada sequência de aminoácidos foi comparada com todas as sequências de aminoácidos de cada genoma e as proteínas foram agrupadas se a pontuação correspondente à similaridade dos pares de sequência fosse maior do que um *z-score* predeterminado, sem considerar o tamanho da região em que houve a correspondência ou o tamanho das sequências, e, com isso, a presença ou ausência de cada grupo de gene em cada genoma foi computada em uma matriz. Em seguida, análises de distância e de parcimônia foram realizadas, assim como foi calculado o valor de *bootstrap*. Árvores filogenéticas foram geradas tanto com os 27 genomas quanto apenas com os 15 genomas de micro-organismos.

Com relação à árvore gerada apenas com os 15 genomas, a árvore mais parcimoniosa foi gerada com um limiar para o *z-score* de 170 como critério para realizar o agrupamento *single linkage* dos grupos de genes homólogos e esta também apresentou, na maioria dos casos, altos valores de *bootstrap*, além de ser consistente com as demais árvores geradas com outros limiares de *z-score* testados. Essa árvore ficou muito similar à gerada com base no gene *16S rRNA*, mas as análises realizadas indicaram que, de maneira geral, houve uma grande dificuldade para explicar as relações entre as proteobactérias, cianobactérias, bactérias Gram-positivas e grupos de bactérias *Deinococcus/Thermus*, o que sugere que esses organismos podem ter passado por significativas transferências horizontais de genes ou de perdas de genes ou que fizeram parte de uma rápida diversificação bacteriana.

Já a árvore gerada a partir dos 27 genomas, apesar de similar à árvore dos 15 genomas e à gerada com base no *16S rRNA*, falhou ao detalhar as relações de diversos organismos, por exemplo, ao unir bactérias do gênero *Mycoplasma* com outras bactérias Gram-positivas e ao unir as proteobactérias em um único ramo. Os autores apontaram

que esses posicionamentos incorretos na árvore podem ter ocorrido devido ao tamanho reduzido dos genomas, visto que a homologia pode ser substancialmente influenciada pela perda de genes de um genoma específico e isso pode ter levado organismos com genomas reduzidos, como as bactérias *Mycoplasma*, a serem colocadas próximas à base da árvore das bactérias. Além disso, o método apresentado pode ser afetado por genes que não estavam presentes no ancestral do grupo ao qual os genomas estão representando. Os autores concluem que uma árvore robusta pode ser reconstruída com base nos grupos de genes homólogos e que esse método sofre menos influência das transferências horizontais de gene do que outros métodos similares baseados em ortólogos.

Korbel et al. (2002), por sua vez, desenvolveram o *web server* SHOT (sigla para o termo, em inglês, *SHared Orthologue and gene-order Tree*), que permite a geração de filogenias com base nas distâncias entre os genomas. A ferramenta permite o uso de duas abordagens distintas para fazer a reconstrução filogenética: uma baseada no conteúdo gênico, na qual a similaridade entre dois genomas é dada pela fração de genes ortólogos compartilhados por eles; e uma abordagem baseada na conservação da ordem dos genes, que só pode ser utilizada para reconstruir árvores para genomas de procariotos, de modo que similaridades são dadas pelo número de pares de genes ortólogos conservados, tal que um par de genes conservados são genes ortólogos que, em dois genomas, formam um par de genes adjacentes na mesma ordem no genoma. Árvores filogenéticas foram reconstruídas utilizando ambas as abordagens e essas foram comparadas com uma árvore reconstruída com base no gene *16S rRNA*.

Tanto a árvore gerada com base no conteúdo gênico quanto a baseada na ordem dos genes apresentaram uma grande similaridade com a árvore gerada com base no gene *16S rRNA*, o que indicou que a conservação na ordem dos genes também pode refletir as distâncias evolucionárias entre as espécies. Apenas na árvore gerada com base na ordem dos genes as β -proteobactérias e γ -proteobactérias foram separadas. Os autores afirmam que a ordem dos genes evolui de maneira mais rápida do que o conteúdo gênico e, por isso, ela pode ser útil para o cálculo de pequenas distâncias evolucionárias, enquanto que o conteúdo gênico é indicado para o cálculo de longas distâncias evolucionárias.

Ye et al. (2013) desenvolveram a ferramenta BlastGraph, que visa à comparação de genomas com base no BLAST e na utilização de grafos. Os resultados obtidos a partir do BLAST podem ser representados por um grafo não direcionado, em que seus vértices representam os aminoácidos ou nucleotídeos das sequências e as arestas representam o

melhor alinhamento para o par de vértices. Duas estratégias são utilizadas sobre esse grafo para se obter informações relevantes: o agrupamento feito com o algoritmo *Markov Cluster* (MLC) e o filtro de arestas. O algoritmo MLC é utilizado para realizar o agrupamento de grafos, em que, primeiramente, gera-se um grafo ponderado a partir do grafo não-direcionado criado com base no *e-value* ou *score* obtidos pelo BLAST e, em seguida, o agrupamento é realizado de modo a se obter subgrafos de acordo com os resultados desse agrupamento. Já o filtro de arestas, que pode ser utilizado como um método complementar para a criação dos subgrafos, é feito com base nos resultados obtidos pelo BLAST (*e-value*, *score*, porcentagem de identidade, etc.). Após o agrupamento, pode ser realizada a reconstrução filogenética, tal que é construída uma matriz referente ao conteúdo gênico dos genomas, indicando a presença ou ausência de cada família de genes para cada genoma. Para realizar o cálculo da distância entre os genomas e, com isso, gerar uma matriz de distância, podem ser utilizados a distância Jaccard ou os métodos *Simple Matching* ou o de Snel, Bork e Huynen (1999), apresentado anteriormente. Com base nessa matriz, podem ser utilizados os algoritmos *neighbor-joining* (NJ) ou UPGMA para a geração da árvore filogenética com base no conteúdo gênico. A ferramenta foi originalmente desenvolvida para analisar dados de genomas de vírus, mas os autores afirmam que ela também pode ser aplicada a grandes genomas de bactérias.

A ferramenta Phylo SI, desenvolvida por Shifman et al. (2014) e que é considerada por eles como mais eficiente para gerar a filogenia de organismos evolutivamente próximos, combina duas abordagens baseadas em genes para realizar a reconstrução filogenética, em que se leva em consideração quanto a ordem dos genes quanto o conteúdo gênico (presença ou ausência dos genes/eventos de ganho ou perda de genes) nos genomas. Para tanto, a ferramenta faz uso de uma nova medida, chamada pelos autores de índice sintênico (do inglês, *synteny index* - SI) entre dois genomas, que visa a medir o quanto os genes ortólogos existentes em duas espécies comparadas estão em seus “espaços naturais”, ou seja, o quanto cada gene ortólogo compartilha o mesmo número de “*k* genes vizinhos” em sua “vizinhança” em ambos os genomas, verificando a conservação da ordem dos genes ao longo da evolução. Após o cálculo desse índice para cada gene e para todas as espécies, obteve-se uma medida de similaridade dada pelo índice SI para as espécies estudadas e, ao calcular o valor médio dessa medida para todo o genoma e ao converter essa medida de similaridade para uma medida de distância, construiu-se uma matriz de distância que permitiu realizar a reconstrução filogenética para os genomas em questão. Após estudos

simulados para estimar o número k ótimo no qual houvesse a maior distinção entre os genomas, de modo que fosse maximizado a separabilidade entre todos os pares de genomas ao mesmo tempo em que não fosse tão pequeno a ponto de ser influenciado por ortologias identificadas incorretamente, a ferramenta foi aplicada a um conjunto de genomas de bactérias.

A árvore filogenética construída, utilizando o algoritmo NJ, apesar de conter diversos ramos correspondendo à filos conhecidos, também apresentou inconsistências para diversas relações taxonômicas conhecidas. Por exemplo, bactérias dos filos *Aquificae* e *Thermotogae* foram erroneamente inseridas no ramos das β -proteobactérias, assim como a β -proteobactéria *Bordetella* foi agrupada erroneamente com γ -proteobactérias. A árvore foi então comparada com outras árvores e observou-se uma alta similaridade entre as árvores baseadas em sequências. Essa alta similaridade também foi observada entre as árvores baseadas em genomas completos, mas menos entre as árvores baseadas em sequências e as baseadas em genomas completos (apesar da árvore gerada pela ferramenta Phylo SI também apresentar alta similaridade com árvores baseadas em sequências).

Os autores concluíram que as similaridades entre árvores geradas com as mesmas abordagens e as divergências entre árvores construídas a partir de abordagens diferentes enfatizam a existência de diferentes processos que ocorrem durante a evolução e que isso sugere a necessidade de uma abordagem mais inclusiva ao considerar a filogenia de procariotos do que considerar apenas a similaridade entre sequências de genes homólogos. Uma limitação do Phylo SI observada foi a de que a aleatoriedade da ordem gênica pode atrapalhar a reconstrução filogenética para as espécies em cujos genomas isso ocorre. A aleatoriedade da ordem dos genes pode ocorrer devido à presença de um grande número de elementos repetidos em genomas que, por sua vez, podem ocorrer especialmente em patógenos intracelulares em que há uma diminuição nas pressões evolutivas para mutações e mudanças genômicas. Entretanto, os autores afirmam que essa limitação também permite que esse método seja útil para identificar tais ocorrências incomuns em genomas.

3.2.3 Considerações finais sobre a revisão não sistemática

Analisando os resultados obtidos pelos trabalhos descritos nessa seção, pode-se notar que há grande dificuldade em se obter um consenso da filogenia de alguns grupos de

bactérias. Isso pode ser notado ao observar o caso das proteobactérias, que é um grupo de bactérias pertencentes ao filo *Proteobacteria* e no qual as *Xanthomonas*, cujos genomas são objeto de estudo do presente projeto, se encontram. No caso dessas bactérias, os trabalhos descritos relatam as dificuldades de se obter uma filogenia que separe as classes de proteobactérias corretamente.

Como mencionado na seção 2.4, a ocorrência de transferências horizontais de genes e a própria dinâmica em que os genomas de bactérias se alteram podem ser algumas das causas para essa dificuldade na reconstrução filogenética, visto que acabam atrapalhando a identificação de características conservadas ao longo da evolução desses micro-organismos.

Outro fato ocorrido na revisão sistemática e que também foi observado na revisão não sistemática foi de que os estudos realizaram, no máximo, uma reconstrução filogenética de espécies diferentes, mas os métodos apresentados por eles não foram aplicados para realizá-la em organismos muito próximos evolutivamente. Um dos estudos inclusive os exclui pela dificuldade em se realizar a reconstrução filogenética desses organismos. Nesse sentido, o presente trabalho é interessante por propor essa reconstrução para organismos muito próximos (como cepas ou patovares de uma mesma espécie ou subespécie) e verificar se a utilização de abordagens similares às que foram apresentadas nesta seção, como a presença ou ausência de genes de famílias de genes homólogos para calcular a distância entre os genomas, podem ser úteis nesse contexto e se elas permitem diferenciá-los uns dos outros.

Além disso, assim como pode ser observado na revisão sistemática, muitos dos trabalhos presentes nessa subseção comparam os resultados obtidos com a filogenia gerada a partir do gene *16S rRNA*, demonstrando o quanto é utilizado como referência para a reconstrução filogenética, apesar de suas limitações já citadas na seção 2.6.

4 Conjunto de ferramentas para comparação de genomas

Este capítulo visa a apresentar um conjunto de ferramentas que foi desenvolvido pelo grupo de pesquisa para realizar a comparação gênica de genomas completos de bactérias. Esse conjunto é dividido em três tipos principais de ferramentas: as que realizam a identificação de genes homólogos, as de comparação dos genomas e as de análise e visualização das redes de genes. Portanto, de acordo com a figura 5, a abordagem que foi adotada neste trabalho para a geração de filogenias é realizada com base nas características de genomas completos na qual se leva em consideração os conteúdos gênicos dos genomas.

As próximas três seções apresentam os três tipos de ferramentas, que estão ilustradas resumidamente na figura 11. Adicionalmente, há a seção 4.4, que apresenta uma descrição de algoritmos desenvolvidos para auxiliar as tarefas realizadas neste trabalho.

Este conjunto de ferramentas, somado ao estudo de caso apresentado nesta dissertação, visam a fornecer uma visão mais robusta e detalhada sobre ferramentas para análise filogenética de organismos filogeneticamente próximos, de forma a auxiliar especialistas do domínio no processo de classificação filogenética. Adicionalmente, pretende-se explorar as vantagens e desvantagens de diferentes abordagens (tanto as já presentes na literatura quanto as propostas neste trabalho).

As ferramentas foram concebidas e implementadas por um conjunto de seis pessoas, a saber: Prof. Dr. Luciano Antonio Digiampietri, orientador do presente projeto; os alunos de iniciação científica Geraldo José dos Santos Júnior e Giovani de Sousa Leite; as alunas de mestrado Vivian Mayumi Yamassaki Pereira, autora desta dissertação, e Priscilla Koch Wagner, cuja dissertação foca na análise de genes exclusivos; e o aluno de doutorado Caio Rafael do Nascimento Santiago. A autora deste trabalho teve maior envolvimento no desenvolvimento das ferramentas apresentadas nas seções 4.1, 4.2 e 4.4.

4.1 Identificação de genes homólogos

A identificação dos genes homólogos pode ser realizada considerando tanto sequências de nucleotídeos quanto de aminoácidos provenientes, respectivamente, de arquivos nos formatos FASTA *fna* e *faa*.

Figura 11 – Resumo das etapas do conjunto de ferramentas para a identificação dos genes homólogos e para a geração das árvores filogenéticas e da rede de genes

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

A partir dessas sequências, dois tipos de alinhamento podem ser executados, por meio da utilização de outros programas já existentes: genes contra genes (utilizando o BLASTP, por exemplo) e genes contra genomas completos (utilizando o TBLASTN, por exemplo). A realização desse último tipo de alinhamento é interessante porque permite a identificação de genes que não foram identificados e anotados e isso poderia atrapalhar a comparação de genomas ao se utilizar apenas a informação dos genes anotados.

O conjunto de ferramentas desenvolvido utiliza como entrada o arquivo de saída, no formato *m8*, do alinhamento do BLAST. Para serem considerados homólogos, dois genes devem ter um alinhamento que satisfaça limiares definidos pelo usuário referentes a sete campos do arquivo *m8*. Também podem ser considerados os valores padrões dos limiares para esses sete campos definidos nas ferramentas, que são:

- *percent identities* (mínimo): 96;
- *alignment percentage* (mínimo): 96;
- *alignment length* (mínimo): 60;
- *number of mismatched positions* (máximo): 20;
- *number of gap positions* (máximo): 5;
- *e-value* (máximo): 10^{-10} ;
- *bit-score* (mínimo): 100.

Duas estratégias são utilizadas para identificar genes homólogos, de modo que, em ambas, a sequência de entrada do alinhamento corresponde a sequências de genes dos genomas, enquanto que as sequências procuradas no banco de dados podem corresponder a sequências de genes ou a sequências completas de DNA dos genomas.

A primeira estratégia consiste no alinhamento dos genes de um genoma contra os demais genes e é utilizada quando se deseja identificar genes presentes no banco de dados que são similares a outros de um genoma específico fornecido como entrada à ferramenta de alinhamento. Nessa estratégia, o gene pertencerá a uma família de genes se o mesmo for alinhado com um gene do genoma específico fornecido como entrada e se esse alinhamento satisfizer os limiares estabelecidos previamente.

Já a segunda estratégia consiste no alinhamento dos genes de todos os genomas contra todos os genomas e é utilizada na comparação de diversos genomas. O gene pertencerá a uma família de genes se ele alinhar com algum gene pertencente a essa família e se esse alinhamento também satisfizer os limiares do alinhamento.

Por fim, oito arquivos no formato *CSV* são produzidos após a identificação dos genes homólogos. Cada arquivo corresponde a uma tabela, em que cada coluna representa um genoma e cada linha representa uma família de genes homólogos. Os valores em cada célula da tabela indicam se o genoma contém ou não um gene da respectiva família. A diferença entre os arquivos, além do conteúdo de cada célula da tabela, é a de que alguns deles fornecem informações específicas, tais como a indicação de famílias exclusivas a um genoma. Os oito diferentes tipos de informação presentes nas células de cada uma dessas tabelas são:

- identificadores dos genes da família gênica presentes no genoma da respectiva coluna da tabela;
- descrição da família gênica presentes no genoma da respectiva coluna da tabela;
- quantidade de genes da família gênica presente no genoma;
- valor normalizado da quantidade de genes da família gênica;
- valor booleano indicando a presença ou ausência de genes da respectiva família de genes;
- valor booleano indicando a presença ou ausência de genes da respectiva família de genes que sejam exclusivos ao respectivo genoma;

- valor booleano indicando os genes “diferentes”, que não estão presentes em todos os genomas;
- valor booleano indicando os genes “diferentes”, mas que não são exclusivos, presentes nos genomas.

4.2 Comparação dos genomas

A comparação de genomas é realizada utilizando os arquivos CSV descritos na seção anterior. Após a comparação, podem ser produzidas árvores filogenéticas.

Para a comparação, duas métricas de distância podem ser utilizadas: a euclidiana e a Manhattan. Os cálculos das distâncias resultam em matrizes quadradas em que cada linha e coluna correspondem a um genoma e cada valor da matriz contém a distância entre os dois genomas da respectiva linha e coluna. Em seguida, um algoritmo de agrupamento hierárquico do pacote *stats* do *software* R pode ser utilizado para produzir filogenias com base nas informações provenientes dessas matrizes de distância.

Além da geração de árvores filogenéticas, as ferramentas também permitem mapear os genomas em um espaço de duas dimensões considerando os dois componentes principais resultantes da Análise de Componentes Principais (do inglês, *Principal Component Analysis* - PCA) (JOLLIFFE, 2002). Tal estratégia mapeia dados multidimensionais (no caso, genomas descritos por milhares de famílias de genes homólogos) em novos dados multidimensionais, tal que cada dimensão, começando da primeira, maximiza a variância dos dados. Dessa forma, as duas dimensões utilizadas para representar o genoma em um espaço de duas dimensões são as que mais representam a variância da informação gênica quando comparando tais genomas. Esse número de coordenadas principais foi escolhido porque permite visualizar a distribuição dos genomas em um espaço bidimensional e pelo fato de que, para os experimentos realizados, as duas primeiras coordenadas principais apresentam uma variação acumulada acima de 70%, o que, segundo Jolliffe (2002), indica o número de coordenadas principais mais adequado para representar o conjunto de dados, cuja variação acumulada deve se encontrar entre 70% e 90%.

4.3 Análise e visualização das redes de genes

Por fim, as ferramentas para visualização e análise das redes de genes permitem a criação de grafos direcionados e não direcionados a partir dos resultados dos alinhamentos, de modo que cada nó no grafo representa um gene e cada aresta representa um alinhamento que satisfaz os limiares estabelecidos. Cada componente conexo no grafo representa uma família de genes homólogos e a função gênica anotada com maior frequência é selecionada para representar a função da respectiva família de genes homólogos. Essas ferramentas são úteis para a compreensão das famílias de genes homólogos identificadas e para auxiliar o usuário na escolha dos limiares para filtrar os alinhamentos.

Diferentes métricas podem ser utilizadas para analisar os grafos. Uma das mais importantes para a análise de famílias de genes homólogos baseada no alinhamento de sequências é o coeficiente de agrupamento, que mede a transitividade das relações (arestas) de um grafo (WATTS; STROGATZ, 1998). Para cada subconjunto de três nós conexos, o coeficiente mede a probabilidade de que esses nós sejam um clique de tamanho três, ou seja, para três nós a , b e c em que haja uma aresta entre os nós a e b e entre b e c , o coeficiente de agrupamento calculará a probabilidade de que haja uma aresta ligando os nós a e c . Dessa forma, os limiares definidos para filtrar os alinhamentos das sequências devem proporcionar altos valores de coeficientes de agrupamento na correspondente rede de genes homólogos, visto que, se os limiares indicam que os nós a e b são homólogos assim como b e c , é interessante que exista uma aresta entre a e c também, já que eles são considerados homólogos por pertencerem ao mesmo componente conexo.

Outras métricas calculadas e que também podem ser utilizadas são o cálculo do número de componentes conexos, distribuição do número de genes por componente conexo, distribuição do número de genomas com genes em cada componente conexo e o grau de distribuição dos genes na rede gênica.

Além disso, uma ferramenta para visualização da rede de genes também foi desenvolvida pelo grupo de pesquisa. Ela permite a geração de uma figura com os componentes conexos selecionados pelo usuário. Essa ferramenta faz uso de um algoritmo *force-directed* para aproximar os nós conectados e distanciar os que não estão conectados.

4.4 Ferramentas complementares

Além das ferramentas descritas nas seções anteriores, outras foram desenvolvidas para complementá-las ou auxiliar em outras etapas dos experimentos realizados neste trabalho. A autora deste trabalho e os alunos de iniciação científica estiveram diretamente envolvidos no desenvolvimento destas ferramentas complementares.

4.4.1 Seleção de genes em um genoma

Um algoritmo para seleção aleatória de genes de famílias de genes homólogos presentes em genomas foi implementado para ser utilizado no experimento que faz a simulação de genomas incompletos.

Para tanto, ele recebeu como entrada a tabela com as quantidades de genes de cada família de genes existentes nos genomas e que foi gerada pelas ferramentas de identificação de genes homólogos descritas anteriormente.

Nos experimentos realizados neste trabalho, o algoritmo fez a seleção de genes de apenas um genoma por vez e a proporção de genes que deveria ser selecionada iniciou-se em 10% e foi aumentando de 10% em 10% até alcançar 90% dos genes presentes no genoma. Por conta da aleatoriedade na seleção, os testes foram repetidos 10 vezes para cada genoma.

Para cada genoma, o algoritmo inicialmente calcula o número total de genes de famílias de genes homólogos presente no genoma e o número de genes que devem ser selecionados para se obter a proporção testada no momento e também zera a coluna contendo as quantidades de genes do respectivo genoma. Em seguida, os genes são selecionados aleatoriamente até que se obtenha o número correspondente à proporção, de modo que a quantidade de genes é atualizada na tabela conforme os genes são selecionados.

Importante mencionar que o algoritmo gera um número aleatório dentre o número total de famílias de genes existentes na tabela e, para cada número selecionado, o algoritmo verifica se a linha da tabela correspondente a esse número se refere a uma família de genes homólogos que contém genes que não estavam originalmente presentes no genoma ou se a quantidade de genes da família de genes homólogos, originalmente presente no genoma, não é excedida ao selecionar mais um gene dessa família. Caso uma dessas condições ocorra, o

gene não é selecionado e o algoritmo gera outro número aleatório para selecionar outro gene.

Por fim, a tabela com a quantidade de genes de cada família de genes nos genomas (contendo as novas quantidades referentes ao genes selecionados na coluna do genoma testado) é então fornecida como entrada para o algoritmo que gera uma matriz de distância dos genomas, calculada com base na distância euclidiana.

4.4.2 Comparação de filogenias

Os algoritmos descritos nessa subseção foram utilizados para comparar filogenias no formato *Newick*. Para isso, duas filogenias são comparadas ao analisar os diferentes grupos formados por genomas presentes na árvore.

Antes da comparação em si, o algoritmo primeiramente verifica se as árvores possuem os comprimentos dos ramos e, caso os possuam, eles são retirados de maneira que apenas as topologias sejam comparadas.

A comparação é realizada de forma que cada grupo formado após um nó interno é comparado com o da outra árvore, sendo que os grupos originados de nós internos mais próximos dos nós terminais são comparados primeiro. Como cada grupo formado após um nó interno é delimitado por parênteses, o algoritmo procura por esses sinais para encontrar os grupos e compara os genomas presentes em cada grupo com os existentes na outra filogenia, de modo que a ordem em que os genomas se encontram no grupo não é considerada; apenas se analisa se todos os genomas presentes em um grupo também estão presentes no grupo da outra filogenia. Após encontrar um grupo e realizar a sua comparação com o grupo observado na outra filogenia, os parênteses são retirados e os genomas passam a pertencer ao grupo formado pelo nó interior anterior, de modo que seja possível comparar os diferentes agrupamentos realizados ao gerar a filogenia.

Ao final da comparação, quando todos os grupos foram comparados, obtém-se o número de grupos diferentes, por apresentarem genomas a mais ou a menos, com relação aos grupos de genomas presentes na árvore comparada, assim como também é obtido o número de grupos que contém exatamente os mesmos genomas dos grupos com os quais foram comparados na outra árvore.

Para exemplificar, podemos considerar uma filogenia no formato *Newick* descrita como $((\mathbf{A},\mathbf{B},\mathbf{C}),\mathbf{D})$; comparada com a filogenia $((\mathbf{A},\mathbf{B}),\mathbf{C},\mathbf{D})$. Inicialmente, o algoritmo compara o grupo composto pelos genomas A, B e C com o grupo da outra filogenia que contém apenas os genomas A e B, visto que esses grupos foram formados por um nó interno mais próximo dos nós terminais. Como falta um genoma nesse grupo, um “erro” é computado. Após a comparação desse grupo, resta apenas comparar o grupo formado pelo nó interno anterior, composto pelos genomas A, B, C e D, que é o mesmo nas duas filogenias, de modo que um “acerto” é computado. Como não há mais nenhum grupo a ser comparado, o algoritmo retorna o número de “erros” e “acertos” encontrados. Nesse exemplo, como houve um “erro”, significa que as duas filogenias comparadas não possuem a mesma topologia.

5 Resultados e discussão

Este capítulo visa a apresentar os experimentos realizados com os genomas de *Xanthomonas*. Primeiramente, no entanto, é apresentada na seção 5.1 uma descrição do conjunto de dados utilizado nos experimentos, assim como das famílias de genes homólogos identificadas nos genomas. A seção 5.2, por sua vez, apresenta as descrições e os resultados dos experimentos realizados. Por fim, a seção 5.3 encerra o capítulo com uma comparação dos resultados obtidos nos experimentos.

5.1 Conjunto de dados

Para todos os experimentos apresentados neste capítulo, foram utilizados os 15 genomas completos de bactérias do gênero *Xanthomonas* mencionados na seção 1.2. A tabela 2 apresenta as informações sobre esses 15 genomas.

Tabela 2 – Informações dos genomas utilizados nos experimentos referentes à abreviação do nome do genoma utilizado, seu tamanho, número de genes e a cor com a qual os seus genes foram representados nas redes de genes geradas pela ferramenta de visualização das redes.

Genoma	Abreviação	Tamanho cromossomo	# de genes	Cor
<i>Xanthomonas albilineans</i> GPE PC73	Xalbilineans	3768695	3208	●
<i>Xanthomonas axonopodis</i> pv. <i>citri</i> str. 306	Xacitri	5175554	4427	●
<i>Xanthomonas axonopodis</i> pv. <i>citrumelo</i> F1	Xacitrumelo	4967469	4181	●
<i>Xanthomonas axonopodis</i> Xac29-1	Xac29-1	5153455	4403	●
<i>Xanthomonas campestris</i> pv. <i>campestris</i> str. 8004	Xcc8004	5148708	4271	●
<i>Xanthomonas campestris</i> pv. <i>campestris</i> str. ATCC 33913	XccATCC33913	5076188	4179	●
<i>Xanthomonas campestris</i> pv. <i>campestris</i> str. B100	XccB100	5079002	4466	●
<i>Xanthomonas campestris</i> pv. <i>raphani</i> 756C	Xraphani	4941214	4516	●
<i>Xanthomonas campestris</i> pv. <i>vesicatoria</i> str. 85-10	Xvesicatoria	5178466	4726	●
<i>Xanthomonas citri</i> subsp. <i>citri</i> Aw12879	Xccitri	5321499	4760	●
<i>Xanthomonas fuscans</i> subsp. <i>fuscans</i>	Xff	4981995	4083	●
<i>Xanthomonas oryzae</i> pv. <i>oryzae</i> KACC 10331	XooKACC10331	4941439	4065	●
<i>Xanthomonas oryzae</i> pv. <i>oryzae</i> MAFF 311018	XooMAFF311018	4940217	4372	●
<i>Xanthomonas oryzae</i> pv. <i>oryzae</i> PXO99A	XooPXO99A	5240075	4988	●
<i>Xanthomonas oryzae</i> pv. <i>oryzicola</i> BLS256	Xooryzicola	4831739	4474	●

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

Com esses dados, foram utilizadas as ferramentas descritas na seção 4.1 para identificação dos genes homólogos, de modo que os alinhamentos foram realizados com a abordagem de genes contra genes, comparando todos os genomas contra todos os genomas. Os valores utilizados para os limiares dos alinhamentos foram os padrões da ferramenta, também apresentados na seção 4.1, que foram escolhidos com base em experimentos realizados previamente de modo que se maximizasse o coeficiente de agrupamento.

A partir das 65.119 sequências de aminoácidos do conjunto de dados (genes), foram obtidos 206.127 alinhamentos após a aplicação dos filtros com base nos limiares definidos. Destes, 50.214 genes alinharam ao menos com um outro gene e 10.090 grupos foram criados com dois ou mais genes. Os tamanhos mais frequentemente encontrados para os grupos foram com quatro genes (38,51% do total), dois genes (17,89%), três genes (16,82%) e cinco genes (9,87%). Já o maior grupo obtido continha 270 genes e correspondia à família de genes *ISX08 transposases*.

Além disso, observou-se que apenas 33 grupos de genes homólogos possuíam genes de todos os genomas (o que corresponde à 0,33% do total de grupos) e 459 grupos (4,55%) possuíam genes de 14 dos 15 genomas de *Xanthomonas*. Os grupos obtidos com maior frequência foram aqueles com genes de apenas quatro genomas (38,97% do total), dois genomas (17,83%) e com três genomas (17,22%). Isso pode ter ocorrido devido à presença do genoma da *Xanthomonas albilineans GPE PC73* no conjunto de dados, que é conhecido na literatura como sendo diferente dos demais genomas utilizados. Com isso, o número de genes compartilhados por todos os genomas pode ter sido menor pelo fato da *Xanthomonas albilineans GPE PC73* não compartilhar tantos genes quanto as demais.

A figura 12 apresenta o grafo com os genes homólogos, de forma que cada gene foi colorido de acordo com as cores dos genomas a que pertencem e que estão descritas na tabela 2. Nessa figura, há 10.090 componentes conexos, que correspondem às famílias de genes homólogos, e 50.214 nós, que correspondem aos genes. Os genes que não foram alinhados com outros genes não estão presentes nessa figura. A figura 13, por sua vez, apresenta uma região de destaque na rede de genes apresentada na figura 12.

Além disso, a partir dos resultados dos alinhamentos, as tabelas descritas na seção 4.1 foram construídas. Para comparar as filogenias obtidas nos experimentos, foi utilizada a tabela que indica o número de genes de cada família de genes homólogos presentes em cada um dos genomas, de modo que uma matriz de distância, calculada com base na distância euclidiana, foi criada e, utilizando o método de agrupamento hierárquico *complete linkage*, uma árvore filogenética de referência foi gerada. Essa filogenia de referência pode ser visualizada na figura 14.

Além dessa filogenia de referência, também foram criadas filogenias, utilizando a ferramenta PhyML, a partir de alinhamentos múltiplos tendo como base a região *16S rRNA*. Para essa região, foram geradas duas filogenias: uma incluindo o genoma da *Xanthomonas albilineans* e outra sem ele. A figura 15 apresenta a filogenia com a *Xanthomonas albilineans*.

Figura 12 – Rede de famílias de genes homólogos

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

Figura 13 – Destaque da rede de famílias de genes homólogos. Por meio dela, é possível notar: um componente densamente povoado, contendo diversos genes (A); um componente com baixo coeficiente de agrupamento (B); um componente composto por um único gene de cada um dos 15 genomas (C); e um componente com apenas quatro genes (D)

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

Figura 14 – Filogenia de referência gerada com base nos genes de famílias de genes homólogos

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

Nota-se que esse genoma ficou separado dos demais. Entretanto, a região *16S rRNA* não foi capaz de distinguir as três *Xanthomonas oryzae* pv. *oryzae* e as três *Xanthomonas citri*. Além disso, também não foi possível distinguir os genomas de *Xanthomonas axonopodis* pv. *citrumelo* F1, *Xanthomonas campestris* pv. *vesicatoria* str. 85-10 e a *Xanthomonas fuscans* subsp. *fuscans*.

A figura 16, por sua vez, apresenta a filogenia da região *16S rRNA* sem o genoma da *Xanthomonas albilineans*. Nela, é possível observar que houve uma mudança na ordem dos grupos formados, de modo que as *Xanthomonas campestris* passaram a descender de um ancestral mais recente do que o que foi observado na filogenia da região *16S rRNA* na qual o genoma da *Xanthomonas albilineans* estava incluído. Adicionalmente, assim como ocorreu na filogenia anterior, não foram diferenciadas as *Xanthomonas oryzae* pv. *oryzae* e as três *Xanthomonas citri*, do mesmo modo como não foi feita a distinção das *Xanthomonas axonopodis* pv. *citrumelo* F1, *Xanthomonas campestris* pv. *vesicatoria* str. 85-10 e a *Xanthomonas fuscans* subsp. *fuscans*.

5.2 Experimentos

As próximas subseções apresentam os quatro experimentos realizados com os genomas de *Xanthomonas*, de modo que:

- a subseção 5.2.1 apresenta o experimento #1, no qual as filogenias foram reconstruídas com base nas distâncias calculadas a partir da ausência ou presença de genes das famílias de genes homólogos em cada genoma;
- a subseção 5.2.2 descreve e apresenta os resultados do experimento #2, no qual foi utilizada a ferramenta GGDC para calcular a distância entre os genomas e, com isso, gerar a filogenia. Os valores estimados de DDH e as probabilidades de que o DDH fosse maior que 70% também foram utilizados para geração de filogenias;
- a subseção 5.2.3 apresenta o experimento #3, em que foram identificados os genes ortólogos presentes em todos os genomas; a partir deles, alinhamentos múltiplos foram realizados e a ferramenta PhyML foi utilizada para gerar a filogenia;
- por último, a subseção 5.2.4 apresenta a descrição e os resultados do experimento #4, no qual foi realizada uma seleção aleatória dos genes a serem considerados em cada genoma para se calcular as distâncias e, com elas, gerar as filogenias.

Figura 15 – Filogenia gerada com base na região *16S rRNA* e incluindo o genoma da *Xanthomonas albilineans*, que foi considerado distante dos demais genomas. Também pode-se observar que não foi possível diferenciar diversos genomas nessa filogenia

Figura 16 – Filogenia gerada com base na região *16S rRNA* sem o genoma da *Xanthomonas albilineans*. Nessa filogenia também não houve a distinção de alguns genomas

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

5.2.1 Experimento #1 - reconstrução filogenética com base na presença ou ausência de genes das famílias de genes homólogos

Nesse experimento, objetivou-se gerar uma filogenia das *Xanthomonas* com base nas famílias de genes homólogos encontradas a partir da utilização das ferramentas descritas na seção 4.1.

Nesse experimento, utilizou-se a tabela com valores binários para indicar a presença ou não de genes de cada grupo de genes homólogos em cada um dos genomas e, a partir dessa tabela, uma matriz de distância dos genomas foi calculada utilizando como medida a distância euclidiana, tal que a distância $d_{x,y}$ entre o genoma X e o genoma Y foi dada pela equação 3:

$$d_{x,y} = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (3)$$

onde n indica o número de famílias de genes e x_i e y_i são os valores binários que indicam se os genomas X e Y , respectivamente, possuíam genes da família de genes i (se igual a um, significa que o genoma possui os genes; se for igual a zero, o genoma em questão não possui genes da família de genes i).

A matriz de distância resultante, por sua vez, pode ser observada na figura 17, em que os genomas foram agrupados de acordo com suas distâncias.

Figura 17 – Matriz de distância entre os genomas considerando a presença ou ausência de genes das famílias de genes homólogos, de modo que, quanto maior o número de famílias de genes homólogos compartilhadas pelos genomas, menor é a distância entre eles

	XccATCC33913	Xcc8004	XccB100	Xraphani	Xalbilineans	Xooryziola	XooPXO99A	XooMAFF311018	XooKACC10331	Xacitrumelo	Xvesicatoria	Xff	Xccitri	Xacitri	Xac29-1
XccATCC33913	0,00	11,83	21,75	25,63	59,94	71,40	74,97	75,41	74,47	75,51	75,76	72,68	78,61	78,91	78,35
Xcc8004	11,83	0,00	20,17	26,85	60,46	71,83	75,49	75,93	74,99	75,91	75,92	73,08	78,56	79,42	78,87
XccB100	21,75	20,17	0,00	22,36	58,79	70,38	74,11	74,57	73,60	74,55	74,64	71,62	77,35	78,06	77,51
Xraphani	25,63	26,85	22,36	0,00	56,02	68,11	71,95	72,39	71,42	72,34	72,75	69,42	75,77	75,98	75,41
Xalbilineans	59,94	60,46	58,79	56,02	0,00	50,78	55,65	56,39	54,87	57,30	57,89	53,17	61,44	61,65	60,84
Xooryziola	71,40	71,83	70,38	68,11	50,78	0,00	28,28	28,86	29,46	58,14	58,79	54,88	62,50	62,88	62,39
XooPXO99A	74,97	75,49	74,11	71,95	55,65	28,28	0,00	17,46	20,93	62,99	63,56	59,83	66,86	67,10	66,61
XooMAFF311018	75,41	75,93	74,57	72,39	56,39	28,86	17,46	0,00	18,63	63,54	64,04	60,42	67,40	67,58	67,16
XooKACC10331	74,47	74,99	73,60	71,42	54,87	29,46	20,93	18,63	0,00	62,56	63,10	59,41	66,51	66,74	66,26
Xacitrumelo	75,51	75,91	74,55	72,34	57,30	58,14	62,99	63,54	62,56	0,00	14,07	37,76	47,22	47,85	47,61
Xvesicatoria	75,76	75,92	74,64	72,75	57,89	58,79	63,56	64,04	63,10	14,07	0,00	38,21	47,27	48,27	48,01
Xff	72,68	73,08	71,62	69,42	53,17	54,88	59,83	60,42	59,41	37,76	38,21	0,00	34,73	35,38	35,14
Xccitri	78,61	78,56	77,35	75,77	61,44	62,50	66,86	67,40	66,51	47,22	47,27	34,73	0,00	19,18	21,52
Xacitri	78,91	79,42	78,06	75,98	61,65	62,88	67,10	67,58	66,74	47,85	48,27	35,38	19,18	0,00	16,52
Xac29-1	78,35	78,87	77,51	75,41	60,84	62,39	66,61	67,16	66,26	47,61	48,01	35,14	21,52	16,52	0,00

Duas representações gráficas foram criadas para sumarizar o resultado na comparação dos genomas, uma usando o método de agrupamento hierárquico *complete linkage* para criar uma árvore filogenética e outra utilizando os resultados do PCA. Os resultados foram comparados com a filogenia criada com base na região *16S rRNA* e com a filogenia de referência.

A figura 18 apresenta a árvore filogenética gerada utilizando as famílias de genes homólogos identificadas, considerando a presença ou não de todos os grupos de genes homólogos. Por meio dessa estratégia, diferentemente do que foi observado na filogenia da região *16S rRNA*, foi possível distinguir as três *Xanthomonas oryzae pv. oryzae* e as três espécies de *Xanthomonas citri*. Além disso, ela foi capaz de agrupar os quatro patovares de *Xanthomonas campestris* e de identificar que a *Xanthomonas albilineans* está provavelmente mais próxima evolutivamente da *Xanthomonas oryzae* do que de outras espécies.

Figura 18 – Filogenia gerada de acordo com a presença ou ausência de genes nos grupos de homólogos. Nela, é possível observar os três principais grupos de *Xanthomonas* e que a *Xanthomonas albilineans* foi colocada próxima das *Xanthomonas oryzae*

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

Com relação à filogenia de referência, houve uma alteração nos posicionamentos das *Xanthomonas oryzae* e das *Xanthomonas campestris*, de modo que a *Xanthomonas*

albilineans tornou-se mais próxima das *Xanthomonas oryzae* e distante das *Xanthomonas campestris* nessa filogenia gerada com base na ausência ou presença de genes homólogos.

Os resultados obtidos a partir da análise do PCA para a ausência ou presença dos genes podem ser visualizados na figura 19. O primeiro componente principal foi mapeado no eixo x da figura e representa 77,49% da variância. Já o segundo componente foi mapeado no eixo y e representa 59,53% da variância restante. Por meio dessa figura, é possível visualizar três grupos principais de genomas: um composto pelas *Xanthomonas campestris*, outro de *Xanthomonas oryzae* e um contendo principalmente *Xanthomonas axonopodis*. Ela também permite notar que *Xanthomonas albilineans* encontra-se entre as *Xanthomonas campestris* e as *Xanthomonas oryzae*, de modo que ela se torna mais próxima de um dos grupos dependendo da abordagem utilizada para a inferência filogenética.

Figura 19 – Mapeamento dos genomas de acordo com as duas componentes principais, em que é possível observar os três principais grupos de *Xanthomonas* e que a *Xanthomonas albilineans* está localizada entre as *Xanthomonas oryzae* e *Xanthomonas campestris*

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

5.2.2 Experimento #2 - reconstrução filogenética com base na ferramenta GGDC

Esta subseção visa a apresentar e discutir sobre as árvores filogenéticas obtidas a partir da utilização da ferramenta GGDC. Para tanto, foram geradas filogenias para cada um dos três resultados obtidos a partir da utilização de cada uma das três fórmulas de distância implementadas pela ferramenta (ver subseção 2.11.2): uma baseada nas distâncias entre cada um dos genomas; uma com base no valor estimado de DDH; e uma a partir da

probabilidade de que o DDH fosse maior que 70% (o que indica a probabilidade de que os genomas realmente correspondam à mesma espécie).

Como resultado, observou-se que, de modo geral, todas as nove filogenias apresentaram agrupamentos similares. Por exemplo, em todas as filogenias, a *Xanthomonas albilineans* ficou separada dos demais genomas e a *Xanthomonas oryzae* pv. *oryzicola* BLS256 foi agrupada com as demais *Xanthomonas oryzae*, ao contrário do que ocorreu na árvore filogenética de referência e na baseada na região *16S rRNA*. Isso pode ser notado, por exemplo, na filogenia gerada com base na distância calculada pela fórmula 1, presente na figura 20. Também é interessante destacar que, nas três filogenias geradas com base na distância, todos os genomas foram diferenciados uns dos outros.

Figura 20 – Filogenia gerada com base nas distâncias calculadas pela ferramenta GGDC utilizando a fórmula 1, na qual há a distinção de todos os genomas e a separação dos três principais grupos de *Xanthomonas*, além da *Xanthomonas albilineans* ter sido considerada como distante das demais

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

Entretanto, apesar dessas similaridades, também foi possível constatar algumas singularidades nas filogenias geradas com base no valor estimado de DDH e na probabilidade de que o DDH fosse maior que 70%. Nessas filogenias, independentemente da fórmula utilizada, não foi possível distinguir os genomas de *Xanthomonas axonopodis* Xac29-1 e de *Xanthomonas axonopodis* pv. *citri* str. 306. Adicionalmente, nas filogenias cujos valores

utilizados foram calculados com base na fórmula 1, também não foi possível distinguir a *Xanthomonas oryzae* pv. *oryzae* KACC 10331 da *Xanthomonas oryzae* pv. *oryzae* MAFF 311018, conforme pode-se verificar na figura 21.

Figura 21 – Filogenia gerada a partir da utilização da fórmula 1 do GGDC para estimar a probabilidade de que o DDH fosse maior que 70%. Nessa filogenia, é possível observar os três principais grupos de *Xanthomonas*, mas não foi possível distinguir alguns dos genomas

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

O pior caso, no entanto, ocorreu a partir da utilização da fórmula 3 para inferir a probabilidade de que o DDH fosse maior que 70%. Nessa filogenia, não foi possível fazer uma distinção dos genomas de *Xanthomonas oryzae* pv. *oryzae* KACC 10331, *Xanthomonas oryzae* pv. *oryzae* PXO99A e *Xanthomonas oryzae* pv. *oryzae* MAFF 311018; das *Xanthomonas campestris* pv. *campestris* str. ATCC 33913, *Xanthomonas campestris* pv. *campestris* str. B100 e *Xanthomonas campestris* pv. *campestris* str. 8004; e das *Xanthomonas axonopodis* pv. *citri* str. 306, *Xanthomonas axonopodis* Xac29-1 e *Xanthomonas citri* subsp. *citri* Aw12879, conforme pode ser constatado na figura 22. Vale destacar que a ferramenta GGDC foi desenvolvida para estimar se dois organismos são ou não da mesma espécie e não para, especificamente, distinguir organismos de uma mesma espécie.

Figura 22 – Filogenia gerada a partir da utilização da fórmula 3 do GGDC para estimar a probabilidade de que o DDH fosse maior que 70%, em que observa-se os três principais grupos de *Xanthomonas*, mas na qual não foi possível diferenciar nove dos 15 genomas

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

5.2.3 Experimento #3 - reconstrução filogenética com base nos genes ortólogos compartilhadas por todos os genomas

Para o experimento #3, foram considerados apenas os grupos de genes homólogos que possuíam genes de todos os genomas, com o objetivo de identificar se os mesmos poderiam ser utilizados para a reconstrução filogenética e se eles apresentavam um melhor desempenho na diferenciação de organismos muito próximos evolutivamente do que o observado com a região *16S rRNA*.

A tabela 3 apresenta uma lista com as 33 famílias de genes homólogos identificadas e que estavam presentes em todos os genomas. Importante notar que foram considerados apenas os genes homólogos que podem ter sido gerados por meio de eventos de especiação, ou seja, que fossem genes ortólogos. Assim, as três últimas famílias presentes na tabela e que estão em destaque não foram utilizadas, uma vez que seus genes possuíam mais de uma cópia em alguns genomas e, portanto, eram potenciais genes parálogos que podem ter sido gerados por eventos de duplicação. Logo, 30 famílias de genes que possuíam um gene presente apenas uma vez em cada genoma foram utilizadas para realizar o experimento.

Tabela 3 – Lista das famílias de genes homólogos presentes em todos os genomas. Em destaque estão as famílias que apresentavam mais de um gene em um genoma e que não foram utilizadas no experimento

#	Família de genes
1	<i>30S ribosomal protein S1</i>
2	<i>30S ribosomal protein S7</i>
3	<i>30S ribosomal protein S9</i>
4	<i>30S ribosomal protein S11</i>
5	<i>30S ribosomal protein S12</i>
6	<i>50S ribosomal protein L13</i>
7	<i>50S ribosomal protein L14</i>
8	<i>50S ribosomal protein L18</i>
9	<i>50S ribosomal protein L22</i>
10	<i>50S ribosomal protein L35</i>
11	<i>acetyl-CoA carboxylase biotin carboxylase subunit</i>
12	<i>acyl carrier protein</i>
13	<i>adenylosuccinate synthetase</i>
14	<i>ATP-dependent Clp protease proteolytic subunit</i>
15	<i>ATP-dependent protease ATP-binding subunit ClpX</i>
16	<i>cell division protein</i>
17	<i>chemotaxis response regulator</i>
18	<i>glycine cleavage system transcriptional repressor</i>
19	<i>integration host factor subunit alpha</i>
20	<i>NADH-ubiquinone oxidoreductase 20 kda subunit</i>
21	<i>rod shape-determining protein MreB</i>
22	<i>transcription antitermination protein NusG</i>
23	<i>transcription regulator protein</i>
24	<i>transcription termination factor Rho</i>
25	<i>translation initiation factor IF-3</i>
26	<i>twitching motility protein</i>
27	<i>two-component system regulatory protein</i>
28	<i>two-component system regulatory protein (response regulator) required for AvrXa21</i>
29	<i>two-component system response regulatory protein (PilG)</i>
30	<i>type II citrate synthase</i>
31	<i>elongation factor Tu</i>
32	<i>translation initiation factor IF-1</i>
33	<i>nucleoside diphosphate kinase</i>

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

Para cada uma das 30 famílias de genes, foi realizado um alinhamento múltiplo com a ferramenta MUSCLE e, após esse alinhamento, a reconstrução filogenética foi feita com a ferramenta PhyML. Para visualização das árvores reconstruídas, foi utilizado o pacote *ape* do *software* R e duas abordagens diferentes foram utilizadas para gerar as árvores: uma considerando o comprimento e a ordem dos ramos e outra considerando apenas a ordem dos ramos. Isso pode ser observado nas figuras 23 e 24, que apresentam as filogenias geradas com base na família de genes *30S ribosomal protein S1*, de modo que a primeira considera os comprimentos dos ramos, enquanto que a última figura, não.

Figura 23 – Filogenia gerada a partir da família de genes *30S ribosomal protein S1*, considerando o comprimento dos ramos

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

Observando as 30 filogenias geradas considerando os comprimentos dos ramos, pode-se notar que elas também apresentaram um problema similar ao que foi observado na filogenia gerada a partir da região *16S rRNA*: diversos genomas apresentaram sequências de genes tão similares que a ferramenta PhyML não pode distingui-los filogeneticamente. O pior caso ocorreu com as famílias de genes *50S ribosomal protein L35* e *integration host factor subunit alpha*, cujas filogenias apresentaram a mesma topologia que, por sua vez, pode ser observada na figura 25. Nessas filogenias, apenas o genoma da *Xanthomonas*

Figura 24 – Filogenia gerada a partir da família de genes *30S ribosomal protein S1* que não considera o comprimento dos ramos

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

albilineans GPE PC73 foi considerado diferente dos demais, pois estava em um ramo muito comprido, e todos os outros genomas foram considerados idênticos.

Algumas das filogenias até foram capazes de agrupar as *Xanthomonas oryzae* e as *Xanthomonas campestris*, por exemplo, como foi o caso da filogenia gerada a partir da família de genes *acetyl-CoA carboxylase biotin carboxylase subunit*, que pode ser visualizada na figura 26. Entretanto, nenhuma das filogenias foi capaz de distinguir os diferentes patovares dessas espécies, assim como nenhuma filogenia foi capaz de diferenciar todos os genomas estudados.

Além disso, ao analisar o comprimento dos ramos, notou-se que o genoma da *Xanthomonas albilineans* GPE PC73 apresentou o ramo mais comprido em 29 das 30 filogenias, o que pode indicar que esse genoma é o mais diferente dentre os demais genomas do conjunto de dados. A única exceção foi a filogenia gerada com base na família de genes *50S ribosomal protein L22*, ilustrada na figura 27, em que é possível notar apenas a separação das *Xanthomonas oryzae* das demais *Xanthomonas*.

Já ao se observar as filogenias sem levar em consideração o comprimento dos ramos, notou-se que em todas as 30 filogenias geradas houve casos de politomias, em que, a partir

Figura 25 – Filogenia gerada a partir da família de genes *50S ribosomal protein L35*. A filogenia gerada a partir da família de genes *integration host factor subunit alpha* apresentou essa mesma topologia. Nelas, apenas a *Xanthomonas albilineans* foi diferenciada das demais

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

Figura 26 – Filogenia geradas a partir da família de genes *acetyl-CoA carboxylase biotin carboxylase subunit*, na qual houve uma certa separação dos três principais grupos de *Xanthomonas*, mas houve a distinção dos diferentes patovares

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

Figura 27 – Filogenia gerada a partir da família de genes *50S ribosomal protein L22*, na qual houve apenas a separação de dois grupos de *Xanthomonas*

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

de um ancestral em comum, há mais de dois descendentes, visto que a partir da raiz de todas as filogenias, há três linhagens descendentes. Algumas das filogenias até foram capazes de separar as *Xanthomonas oryzae* das *Xanthomonas campestris*, por exemplo, mas não foram capazes de agrupá-las. Isso pode ser observado na filogenia gerada com base na família de genes *50S ribosomal protein L18*, que pode ser visualizada na figura 28, em que as *Xanthomonas oryzae* foram separadas das demais *Xanthomonas*, mas o uso desse gene não foi suficiente para agrupá-las e distingui-las. O mesmo pode ser observado na figura 29, que apresenta a filogenia baseada na família de genes *transcription antitermination protein NusG*, em que houve o agrupamento e distinção das *Xanthomonas oryzae*, mas na qual não foi possível agrupar e distinguir as *Xanthomonas campestris* localizadas na parte inferior da árvore.

Com base no algoritmo implementado para comparar, automaticamente, as topologias das filogenias geradas pelo PhyML, sem considerar o comprimento dos ramos, observou-se que nenhuma das 30 famílias de genes produziu uma filogenia idêntica à árvore filogenética de referência. Também pode-se observar que a topologia mais frequente, presente na figura 30, apareceu apenas duas vezes e foi observada ao utilizar as famílias de genes *50S ribosomal protein L14* e *acyl carrier protein*. Nenhuma das outras famílias

Figura 28 – Filogenia gerada a partir da família de genes *50S ribosomal protein L18*, na qual nota-se que o grupo de *Xanthomonas oryzae* não foi propriamente agrupado

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

Figura 29 – Filogenia gerada a partir da família de genes *transcription antitermination protein NusG*, em que é possível observar que o grupo de *Xanthomonas campestris* não foi propriamente agrupado

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

de genes produziu uma filogenia idêntica a de outras famílias com relação aos grupos e à ordem dos genomas na árvore.

Figura 30 – Filogenia mais frequente, gerada a partir das famílias de genes *50S ribosomal protein L14* e *acyl carrier protein*

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

Por fim, é interessante notar que as diferenças observadas visualmente entre as filogenias considerando ou não o comprimento dos ramos ocorreu porque muitos deles eram de zero ou de apenas 10^{-8} . Desse modo, apesar de ordenar e agrupar os genomas, algumas das filogenias não conseguiram distingui-los pelo fato do comprimento dos ramos ser muito pequeno ou ser igual a zero.

5.2.4 Experimento #4 - reconstrução filogenética com base nas proporções de genes dos genomas

Nesse experimento, foram estudadas as proporções dos genomas que seriam necessárias para se obter uma filogenia idêntica à filogenia de referência. Esse estudo é interessante porque simula a utilização de genomas incompletos, pois apesar do sequenciamento ter se tornado uma tarefa mais rápida e barata do que nas décadas passadas, muitos dos genomas armazenados em bancos de dados ainda estão incompletos devido ao fato de que é mais barato sequenciar genomas incompletos do que genomas completos, visto que os custos para se terminar o sequenciamento de um genoma podem chegar a mais de 95%

do custo total (LAND et al., 2015). Desse modo, esse estudo permitiu verificar se, para as *Xanthomonas* analisadas, é possível gerar uma filogenia que diferencie os genomas e que vá de encontro com a filogenia de referência sem se ter a necessidade do sequenciamento completo de seus genomas.

Para isso, foram realizados experimentos individuais com cada um dos 15 genomas, de modo que, com base na tabela que contém o número de genes de cada família de genes presente em cada genoma (que também foi utilizada para gerar a filogenia de referência), um algoritmo desenvolvido selecionou, aleatoriamente, apenas uma porção desses genes. Essas porções foram testadas no intervalo de 10% a 90%, variando a cada 10%. Assim, o genoma testado continha em sua coluna na tabela apenas os genes selecionados, enquanto os demais genomas continuavam com a mesma quantidade e os mesmos genes que possuíam. Com base nessa nova tabela, uma matriz de distância foi calculada utilizando a distância euclidiana, em que a distância $d_{x,y}$ entre o genoma X e o genoma Y foi dada pela equação 4:

$$d_{x,y} = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (4)$$

onde n indica o número de famílias de genes e x_i e y_i indicam o número de genes que os genomas X e Y , respectivamente, possuíam da família de genes i .

Em seguida, para cada teste realizado, uma filogenia foi gerada utilizando a função *hclust* do *software* R, cujo método padrão para realizar o agrupamento hierárquico é o *complete linkage*, que recebeu como entrada a matriz de distância calculada anteriormente.

Por conta da aleatoriedade na seleção dos genes, o experimento foi repetido 10 vezes para cada genoma. Logo, foram geradas 90 filogenias para cada genoma e 1.350 filogenias no total.

Para comparar as filogenias com a filogenia de referência, calculou-se a correlação tanto da distância entre os genomas quanto da posição relativa de todos os genomas em relação ao genoma incompleto de cada experimento. A correlação $\rho_{x,y}$ entre as posições e entre as distâncias da filogenia de referência X e da filogenia comparada Y foi calculada com base na equação 5:

$$\rho_{x,y} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}} \quad (5)$$

onde $\bar{x}$ e $\bar{y}$ indicam as médias das distâncias (ou posições), respectivamente, da filogenia de referência e da filogenia comparada.

Por fim, para cada porcentagem de genes testada para cada genoma, calculou-se a média e a mediana das correlações das posições e distâncias dos genomas obtidas nas 10 iterações.

Observando essas medidas, notam-se alguns resultados interessantes. Por exemplo, todas as variações nas proporções de genes presentes no genoma da *Xanthomonas albilineans* foram as únicas em que a correlação das posições dos genomas na árvore filogenética foi igual a 1, que é máxima, em todas as 10 iterações realizadas, de modo que a média e a mediana da correlação das posições na filogenia também foram iguais a 1. Além disso, as distâncias entre os genomas também apresentaram uma correlação muito forte. Todos esses dados estão presentes na tabela 4.

Tabela 4 – Médias e medianas de correlação das posições e distâncias dos genomas na filogenia quando variada a porção de genes da *Xanthomonas albilineans*

% genes	Distância		Posição	
	Média	Mediana	Média	Mediana
10%	0,993242552	0,993251876	1	1
20%	0,994747831	0,99475059	1	1
30%	0,9960293	0,996033124	1	1
40%	0,997106852	0,997113936	1	1
50%	0,997998399	0,997993832	1	1
60%	0,998715748	0,998709531	1	1
70%	0,999262919	0,999257113	1	1
80%	0,999660057	0,999656299	1	1
90%	0,99990185	0,999899146	1	1

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

Esses valores podem ter sido obtidos porque os genes presentes na *Xanthomonas albilineans* são diferentes dos presentes nos demais genomas, de modo que mesmo ao considerar somente 10% dos genes, já é possível posicionar a *Xanthomonas albilineans* corretamente na árvore filogenética e obter a mesma topologia da filogenia de referência.

Os resultados obtidos com a variação da proporção dos genes da *Xanthomonas oryzae* pv. *oryzae* KACC 10331 e da *Xanthomonas oryzae* pv. *oryzae* MAFF 311018, por outro lado, obtiveram valores de correlação muito baixos. Para este último genoma, nem mesmo a seleção de 90% dos genes permitiu que a filogenia gerada ficasse próxima da de referência, visto que os valores das médias e medianas da correlação das distâncias e das posições foram baixas, conforme pode ser observado na tabela 5.

Tabela 5 – Médias e medianas de correlação das posições e distâncias dos genomas na filogenia quando variada a porção de genes da *Xanthomonas oryzae pv. oryzae* MAFF 311018

% genes	Distância		Posição	
	Média	Mediana	Média	Mediana
10%	0,007486332	0,007598474	-0,021069382	-0,017857143
20%	0,039653582	0,03893315	0,246071429	0,255357143
30%	0,07354475	0,072666009	0,30599273	0,304963651
40%	0,110142811	0,109389995	0,342857143	0,344642857
50%	0,146379175	0,146372431	0,368928571	0,364285714
60%	0,182729242	0,183794269	0,370714286	0,364285714
70%	0,226327552	0,228693727	0,371071429	0,364285714
80%	0,282585684	0,285444074	0,3675	0,367857143
90%	0,479400624	0,476579425	0,522142857	0,525

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

Já os resultados para a *Xanthomonas oryzae pv. oryzae* KACC 10331 apresentaram correlações ainda mais baixas, valores esses que estão presentes na tabela 6. Todas as correlações, tanto de distância quanto de posição, apresentaram valores muito próximos de 0, até mesmo ao selecionar 90% dos genes para gerar a filogenia. Também é possível notar muitas correlações negativas, indicando que as posições e distâncias da filogenia comparada variam na direção oposta das da filogenia de referência.

Tabela 6 – Médias e medianas de correlação das posições e distâncias dos genomas na filogenia quando variada a porção de genes de *Xanthomonas oryzae pv. oryzae* KACC 10331

% genes	Distância		Posição	
	Média	Mediana	Média	Mediana
10%	-0,196297114	-0,196413334	-0,1725	-0,176785714
20%	-0,178464844	-0,178613833	-0,083630833	-0,082142857
30%	-0,159306353	-0,158553488	-0,034285714	-0,039285714
40%	-0,139996169	-0,140064954	-0,006785714	-0,003571429
50%	-0,117904294	-0,119251615	0,021381146	0,021428571
60%	-0,093469777	-0,093748627	0,024285714	0,019642857
70%	-0,065877271	-0,066523436	0,025714286	0,019642857
80%	-0,029113399	-0,028750329	0,025357143	0,021428571
90%	0,136745146	0,136624001	0,090357143	0,092857143

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

Por fim, a tabela 7 resume alguns resultados obtidos para cada um dos 15 genomas estudados. Nela, estão descritas as proporções de genes necessárias para que as medianas das correlações com a árvore de referência, tanto das distâncias entre os genomas quanto das posições na filogenia, fossem maiores do que 0,7, que é uma correlação considerada alta

(CALLEGARI-JACQUES, 2008). Para a maioria dos experimentos realizados, correlações acima de 0,84 representaram resultados nos quais todas as espécies foram corretamente separadas, apresentando problemas, no máximo, na filogenia relativa às cepas de uma mesma espécie.

Tabela 7 – Proporção de genes necessária para que as medianas das correlações de distância e posição fossem acima de 0,7. Entre parênteses estão os valores obtidos nessas medianas e o hífen indica que não foi obtida nenhuma mediana acima de 0,7 para nenhuma proporção testada

Genoma	Distância	Posição
<i>Xanthomonas albilineans</i> GPE PC73	10% (0,993251876)	10% (1)
<i>Xanthomonas axonopodis</i> pv. <i>citri</i> str. 306	10% (0,935148651)	10% (0,725)
<i>Xanthomonas axonopodis</i> pv. <i>citrumelo</i> F1	10% (0,974635665)	10% (0,710714286)
<i>Xanthomonas axonopodis</i> Xac29-1	10% (0,948049877)	10% (0,725)
<i>Xanthomonas campestris</i> pv. <i>campestris</i> str. 8004	10% (0,931213011)	10% (0,844642857)
<i>Xanthomonas campestris</i> pv. <i>campestris</i> str. ATCC 33913	10% (0,93904613)	10% (0,892857143)
<i>Xanthomonas campestris</i> pv. <i>campestris</i> str. B100	10% (0,956427366)	10% (0,939285714)
<i>Xanthomonas campestris</i> pv. <i>raphani</i> 756C	10% (0,960979889)	10% (0,896428571)
<i>Xanthomonas campestris</i> pv. <i>vesicatoria</i> str. 85-10	10% (0,966782036)	10% (0,769642857)
<i>Xanthomonas citri</i> subsp. <i>citri</i> Aw12879	10% (0,944159569)	10% (0,844642857)
<i>Xanthomonas fuscans</i> subsp. <i>fuscans</i>	10% (0,97041034)	10% (0,8875)
<i>Xanthomonas oryzae</i> pv. <i>oryzae</i> KACC 10331	-	-
<i>Xanthomonas oryzae</i> pv. <i>oryzae</i> MAFF 311018	-	-
<i>Xanthomonas oryzae</i> pv. <i>oryzicola</i> BLS256	10% (0,746841728)	-
<i>Xanthomonas oryzae</i> pv. <i>oryzae</i> PXO99A	20% (0,746968276)	30% (0,769642857)

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

Por meio dessa tabela, é possível notar que a maioria dos genomas de *Xanthomonas* apresentaram uma mediana de correlação alta mesmo ao selecionar apenas 10% de seus genomas para a geração da filogenia. As únicas exceções foram observadas com relação às *Xanthomonas oryzae*. A *Xanthomonas oryzae* pv. *oryzae* PXO99A, apesar de não ter obtido uma mediana acima de 0,7 ao utilizar apenas 10% dos genes, obteve valores de correlação altos quando selecionados mais de 20% (para a correlação das distâncias entre os genomas) e 30% (para a correlação das posições dos genomas nas filogenias) dos genes. Por outro lado, as *Xanthomonas oryzae* pv. *oryzae* KACC 10331 e *Xanthomonas oryzae* pv. *oryzae* MAFF 311018, como já foi mencionado anteriormente, obtiveram correlações muito baixas mesmo com 90% dos genes. Por fim, a *Xanthomonas oryzae* pv. *oryzicola* BLS256, apesar de ter apresentado uma mediana alta para a correlação das distâncias dos genomas, não obteve uma mediana acima de 0,7 para a correlação das posições, mesmo ao utilizar 90% dos genes. Também é interessante citar que a sua mediana para a correlação das posições, ao contrário do que ocorreu nos demais genomas de modo geral, foi decrescendo conforme a proporção de genes selecionados aumentou.

Adicionalmente, as filogenias também foram comparadas visualmente para verificar qual a proporção de genes do genoma necessária para se obter a filogenia de referência já apresentada previamente na figura 14.

Com base nessas análises, constatou-se que as filogenias geradas com todas as variações nas proporções dos genes da *Xanthomonas albilineans* em todas as iterações apresentaram a mesma topologia da árvore filogenética de referência. Essa mesma topologia e o fato de apresentar altos valores de correlação das posições dos genomas e das distâncias entre eles são indícios de que a *Xanthomonas albilineans* apresenta o conteúdo gênico mais distinto das demais e, por isso, mesmo com poucos genes é possível posicioná-la corretamente na árvore filogenética.

Outro fato que pode ser observado nas 10 iterações é de que as filogenias geradas com 60%, 70% e 80% dos genes da *Xanthomonas campestris* pv. *campestris* str. 8004 obtiveram a mesma topologia da filogenia de referência, com exceção da ordem das *Xanthomonas campestris* pv. *campestris* em seu agrupamento. Isso pode ser observado na figura 31, em que a *Xanthomonas campestris* pv. *campestris* str. ATCC 33913 e a *Xanthomonas campestris* pv. *campestris* str. B100 estão mais próximas evolutivamente, quando na árvore de referência a *Xanthomonas campestris* pv. *campestris* str. ATCC 33913 está mais próxima evolutivamente da *Xanthomonas campestris* pv. *campestris* str. 8004. Isso indica que, para esse genoma, é necessário considerar mais do que 80% dos seus genes para que ele se localize na ordem correta da filogenia.

Além disso, também foi elaborada a tabela 8, que indica qual a porcentagem mínima de genes observada que foi necessária para se obter uma filogenia idêntica à árvore filogenética de referência nas 10 iterações realizadas. Nos casos em que houve uma variação da proporção mínima nas 10 iterações realizadas, é apresentado o intervalo das proporções mínimas observadas e um hífen indica os casos em que não houve uma proporção de genes que permitisse obter a topologia da filogenia de referência.

Analisando a tabela, percebe-se que para nove dos 15 genomas, o que corresponde à 60% do total de genomas analisados, não seria necessário fazer uso de todos os seus genes para se obter uma filogenia com topologia idêntica à referência, sendo que para quatro deles é necessário considerar a metade ou menos dos genes que possuem. Isso é interessante porque demonstra que, para esses genomas, não é necessário realizar o sequenciamento completo, cujo custo pode ser muito elevado comparado com o do sequenciamento incompleto, para que eles sejam posicionados corretamente na filogenia.

Figura 31 – Filogenia gerada em um das iterações utilizando 80% dos genes da *Xanthomonas campestris* pv. *campestris* str. 8004, na qual a *Xanthomonas campestris* pv. *campestris* str. ATCC 33913 está mais próxima evolutivamente de uma *Xanthomonas* distinta a que estava na árvore de referência

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

Tabela 8 – Proporção mínima de número de genes necessária para que as filogenias geradas fossem idênticas à filogenia de referência

Genoma	Proporção mínima
<i>Xanthomonas albilineans</i> GPE PC73	10%
<i>Xanthomonas axonopodis</i> pv. <i>citri</i> str. 306	80% - 90%
<i>Xanthomonas axonopodis</i> pv. <i>citrumelo</i> F1	50%
<i>Xanthomonas axonopodis</i> Xac29-1	70% - 80%
<i>Xanthomonas campestris</i> pv. <i>campestris</i> str. 8004	80%
<i>Xanthomonas campestris</i> pv. <i>campestris</i> str. ATCC 33913	70% - 80%
<i>Xanthomonas campestris</i> pv. <i>campestris</i> str. B100	60% - 80%
<i>Xanthomonas campestris</i> pv. <i>raphani</i> 756C	40%
<i>Xanthomonas campestris</i> pv. <i>vesicatoria</i> str. 85-10	-
<i>Xanthomonas citri</i> subsp. <i>citri</i> Aw12879	-
<i>Xanthomonas fuscans</i> subsp. <i>fuscans</i>	30%
<i>Xanthomonas oryzae</i> pv. <i>oryzae</i> KACC 10331	-
<i>Xanthomonas oryzae</i> pv. <i>oryzae</i> MAFF 311018	-
<i>Xanthomonas oryzae</i> pv. <i>oryzicola</i> BLS256	-
<i>Xanthomonas oryzae</i> pv. <i>oryzae</i> PXO99A	-

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

Por outro lado, não foi possível obter uma filogenia com a topologia idêntica a da filogenia de referência para nenhuma das variações das proporções dos genes das quatro *Xanthomonas oryzae* (*Xanthomonas oryzae* pv. *oryzae* KACC 10331, *Xanthomonas oryzae* pv. *oryzae* MAFF 311018, *Xanthomonas oryzae* pv. *oryzicola* BLS256 e *Xanthomonas oryzae* pv. *oryzae* PXO99A) em nenhuma das 10 iterações.

Além desses genomas, as filogenias com variações dos genomas da *Xanthomonas citri* subsp. *citri* Aw12879 e da *Xanthomonas campestris* pv. *vesicatoria* str. 85-10, apesar

de terem apresentado filogenias com altas medianas de correlação das posições na árvore e das distâncias dos genomas a partir da utilização de apenas 10% de seus conteúdos gênicos, também não apresentaram uma topologia idêntica a da árvore filogenética de referência.

Comparando os valores da tabela 7 com os presentes na tabela 8, é possível observar que, apesar de grande parte dos genomas apresentar uma mediana da correlação acima de 0,7, mesmo a partir da utilização de apenas 10% de seus genes, é necessário mais do que essa proporção para que se obtenha uma topologia igual a da árvore de referência para a maioria dos genomas analisados. Ao analisar as correlações obtidas para as proporções mínimas necessárias para que se gerasse uma topologia idêntica à da filogenia de referência, notou-se que as medianas das correlações de distância e de posição dos genomas apresentaram valores muito elevados, maiores do que 0,97, sendo que, para as proporções mínimas das *Xanthomonas axonopodis* e da *Xanthomonas campestris* pv. *campestris* str. 8004, esses valores ficaram acima de 0,99, o que demonstra que quase a totalidade de seus genes é necessária para que se esses genomas se posicionem corretamente na árvore filogenética. Portanto, de modo geral, a utilização de poucos genes, apesar de ter permitido separar diferentes espécies e obter correlações altas, não foi suficiente para separar diferentes cepas de espécies, de forma que, para a maioria dos genomas, foram necessárias proporções maiores (mas não a totalidade) de genes para diferenciar os organismos muito próximos evolutivamente e obter uma árvore idêntica à de referência.

Por fim, também observou-se as proporções mínimas de genes necessárias para que os três principais grupos de *Xanthomonas* (um composto por *Xanthomonas oryzae*, outro por *Xanthomonas campestris* e outro com as demais *Xanthomonas*) fossem separados nas árvores filogenéticas geradas em cada uma das iterações. A tabela 9 apresenta os resultados obtidos com essa análise.

Observando esses resultados, é possível notar que, para muitos genomas, apesar de não ter sido possível distinguir todos os diferentes patovares e subespécies, foi possível distinguir os três principais grupos de *Xanthomonas* com proporções menores (e iguais ou próximas às apresentadas na tabela 7, indicando que elas apresentam altas correlações de distâncias e posições na filogenia) do que o necessário para distinguir todos os genomas; apenas com três *Xanthomonas oryzae* não foi possível obter uma filogenia na qual os três principais grupos de *Xanthomonas* estivessem diferenciados. Além disso, também foi possível observar que, apesar da maioria das filogenias distinguir os três grupos, a ordem de proximidade entre eles e a localização da *Xanthomonas albilineans* GPE PC73 variaram

Tabela 9 – Proporção mínima de genes necessária para que os genomas fossem separados nos três principais grupos de *Xanthomonas* nas filogenias geradas

Genoma	Proporção mínima
<i>Xanthomonas albilineans</i> GPE PC73	10%
<i>Xanthomonas axonopodis</i> pv. citri str. 306	50%
<i>Xanthomonas axonopodis</i> pv. citrumelo F1	50%
<i>Xanthomonas axonopodis</i> Xac29-1	20% - 30%
<i>Xanthomonas campestris</i> pv. campestris str. 8004	10%
<i>Xanthomonas campestris</i> pv. campestris str. ATCC 33913	10%
<i>Xanthomonas campestris</i> pv. campestris str. B100	10%
<i>Xanthomonas campestris</i> pv. raphani 756C	10%
<i>Xanthomonas campestris</i> pv. vesicatoria str. 85-10	20%
<i>Xanthomonas citri</i> subsp. citri Aw12879	20%
<i>Xanthomonas fuscans</i> subsp. fuscans	30%
<i>Xanthomonas oryzae</i> pv. oryzae KACC 10331	-
<i>Xanthomonas oryzae</i> pv. oryzae MAFF 311018	-
<i>Xanthomonas oryzae</i> pv. oryzicola BLS256	-
<i>Xanthomonas oryzae</i> pv. oryzae PXO99A	30% - 40%

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

dependendo do genoma cuja proporção de genes foi variada, conforme é possível observar nas figuras 32 e 33, por exemplo.

Figura 32 – Filogenia gerada em um das iterações utilizando 10% dos genes da *Xanthomonas campestris* pv. campestris str. 8004, na qual o grupo de *Xanthomonas oryzae* é o mais distante

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

5.3 Considerações finais

Este capítulo teve como objetivo apresentar os resultados de quatro experimentos realizados com 15 genomas completos de bactérias do gênero *Xanthomonas*.

De modo geral, pode-se dizer que foi possível obter uma árvore filogenética com três principais grupos de genomas (um grupo com genomas de *Xanthomonas oryzae*,

Figura 33 – Filogenia gerada em um das iterações utilizando 30% dos genes da *Xanthomonas oryzae* pv. *oryzae* PXO99A, na qual o grupo de *Xanthomonas campestris* é o mais distante

Fonte: Vivian Mayumi Yamassaki Pereira, 2017

outro com genomas de *Xanthomonas campestris* e o outro composto principalmente pelas *Xanthomonas axonopodis*) nas filogenias geradas com base na ausência ou presença de genes das famílias de homólogos, diferentes proporções de genes homólogos nos genomas e quando foi utilizada as distâncias dos genomas e os valores e probabilidades do DDH calculados pela ferramenta GGDC. A exceção ocorreu nas filogenias geradas com base nos alinhamentos múltiplos dos genes de famílias presentes em todos os genomas, visto que, nesses casos, houve a definição de dois dos grupos de genomas, mas o terceiro não foi propriamente agrupado.

Com relação a esse experimento, concluiu-se que a utilização, separadamente, dos genes ortólogos presentes em todos os genomas não foi adequada para a distinção dos genomas evolutivamente próximos e nem para distinguir os três principais grupos de *Xanthomonas* estudados, visto que em nenhuma filogenia gerada a partir desses genes houve, concomitantemente, uma clara separação das *Xanthomonas oryzae*, *Xanthomonas campestris* e do grupo composto principalmente de *Xanthomonas axonopodis*. Adicionalmente, com base nesses genes ortólogos compartilhados, apesar de considerar o ramo de *Xanthomonas albilineans* como o mais comprido, esse genoma foi agrupado com as demais *Xanthomonas*, apesar de outras abordagens indicarem que ele é evolutivamente distante dos demais genomas analisados neste trabalho.

Já o experimento que utilizou diferentes proporções de genes presentes em cada genoma para geração da filogenia de modo a simular o uso de genomas incompletos indicou

que, para muitos dos genomas estudados, não é necessário considerar todos os genes para que se obtenha uma filogenia com topologia igual a da árvore de referência e com altos valores de correlação das posições e distâncias dos genomas com esta árvore. As exceções notadas foram, principalmente, as *Xanthomonas oryzae*, em que não foi possível gerar filogenias correspondentes à filogenia de referência com nenhuma variação das proporções dos genes utilizados no experimento, indicando que, para esses genomas, é necessário mais do que 90% de seu conteúdo gênico para que se obtenha uma árvore filogenética idêntica a de referência. Ainda com relação a esse experimento, também pode-se observar que, para a maioria dos genomas, seria possível distinguir os três principais grupos das *Xanthomonas* estudadas mesmo considerando apenas pequenas proporções de seus genes, o que pode indicar que as famílias de genes homólogos permitem distinguir genomas de grupos diferentes porque o número de genes homólogos compartilhados por genomas pertencentes a um mesmo grupo é maior do que o número compartilhado por genomas de grupos diferentes. Entretanto, para distinguir não só os diferentes grupos como também os genomas pertencentes a esses grupos, seriam necessárias proporções maiores dos genomas, visto que apenas com pequenas proporções há dificuldade para diferenciar e determinar a proximidade do genoma, cuja proporção de genes foi variada, com os demais dentro de seu grupo.

Além disso, é interessante notar que, por conta da proximidade evolutiva de alguns dos genomas, nas filogenias geradas com base no valor estimado de DDH e em sua probabilidade, nos genes compartilhados por todos os genomas e até mesmo nas filogenias com base na região *16S rRNA*, amplamente utilizado como referência, não foi possível distinguir todos os genomas analisados nos experimentos. Isso só foi possível a partir da análise da ausência ou presença de genes das famílias de genes nos genomas, do número de genes das famílias de genes homólogos presentes em cada genoma e da distância entre genomas calculada pela ferramenta GGDC.

Outro aspecto importante que deve ser salientado é quanto ao genoma da *Xanthomonas albilineans*. Apesar das filogenias geradas com base na região *16S rRNA* e no valor estimado de DDH não conseguirem fazer a distinção de genomas muito próximos evolutivamente, elas demonstram claramente que a *Xanthomonas albilineans* está mais distante dos demais genomas do conjunto de dados analisado. Além disso, ao analisar estudos realizados por outros autores como [Pieretti et al. \(2009\)](#), [Sharma e Patil \(2010\)](#) e [Rodriguez-R et al. \(2012\)](#), que reconstruíram as filogenias com base em outros genes, pode-se concluir que

a *Xanthomonas albilineans* está mais distante das demais *Xanthomonas* utilizadas neste trabalho, tal como é apresentada na filogenia gerada com base na região *16S rRNA*.

A filogenia utilizada como referência neste trabalho e que foi construída com base no número de genes das famílias de genes homólogos existentes nos genomas, apesar de distinguir os genomas de *Xanthomonas* dos três grupos principais encontrados, mesmo os mais próximos, acabou por inserir a *Xanthomonas albilineans* no grupo das *Xanthomonas campestris*. O mesmo ocorreu com as filogenias baseadas nos genes ortólogos compartilhados por todos os genomas, pois apesar de representá-la com o ramo de maior comprimento, esse genoma também foi agrupado com os demais. As filogenias geradas com variadas proporções de genes da *Xanthomonas albilineans*, apesar de demonstrarem que a utilização de uma pequena proporção do genoma já seria suficiente para obter altas correlações de posições e de distâncias entre os genomas, o que poderia indicar que esse genoma é tão diferente dos demais que não seria necessário muitos genes para fazer essa diferenciação, também agruparam a *Xanthomonas albilineans* com as *Xanthomonas campestris*.

Desse modo, pode-se concluir que a reconstrução filogenética baseada nas famílias de genes homólogos nos genomas de *Xanthomonas* pode ser útil para a identificação e diferenciação de genomas muito próximos evolutivamente, enquanto que os métodos amplamente utilizados, como o DDH e a região *16S rRNA*, são importantes para distinguir genomas de espécies diferentes e evolutivamente mais distantes de *Xanthomonas*, como foi o caso da *Xanthomonas albilineans*.

Além disso, é importante salientar que, apesar dos experimentos realizados utilizarem um conjunto de genomas de bactérias apenas do gênero *Xanthomonas*, as ferramentas desenvolvidas e as diferentes abordagens adotadas para a reconstrução e comparação de filogenias, apresentadas nesta dissertação, podem ser utilizadas para realizar a reconstrução filogenética de outros genomas de procariotos. Os resultados obtidos com os genomas de *Xanthomonas* fornecem indícios de que essas ferramentas e abordagens podem ser úteis principalmente para a reconstruções filogenéticas de genomas de procariotos muito próximos evolutivamente, como patovares de uma mesma espécie, feitas com base nas famílias de genes homólogos identificadas. Esses resultados também indicam que algumas das abordagens para geração de filogenias com base nas características de genomas completos apresentadas na subseção 3.2.2 de trabalhos correlatos e que são similares às adotadas neste trabalho, como a utilização dos genes homólogos para inferir a filogenia ou de grafos para encontrar as famílias de genes homólogos, também podem ser úteis para

a geração de filogenias de genomas evolutivamente próximos, apesar delas motivarem a reconstrução filogenética de, no máximo, espécies.

É relevante mencionar que o algoritmo e a estratégia para seleção de genes de genomas também podem ser utilizados para outros genomas de procariotos, uma vez que eles são um indicativo para a necessidade ou não de se sequenciar os genomas completamente dependendo do estudo desejado (por exemplo, para se obter uma filogenia com topologia idêntica a de uma determinada filogenia utilizada como referência), o que pode diminuir os custos desse estudo caso não seja necessário realizar o sequenciamento completo. Outra aplicabilidade que as ferramentas podem ter é auxiliar a definição de filogenias de organismos cujos genomas não foram sequenciados anteriormente. Nesse caso, as ferramentas permitem identificar os genomas com os quais o genoma desse organismo é mais similar e podem, por exemplo, ajudar a verificar se esse organismo pertence a uma nova espécie.

Por fim, é interessante citar que a estratégia utilizada neste trabalho para a identificação das famílias de genes homólogos e que considera o alinhamento de genes com o DNA de genomas completos, pode minimizar os problemas que podem ocorrer quando a análise filogenética é realizada considerando apenas um gene específico, como os que ocorrem quando há genes que não foram identificados ou anotados ou que foram anotados incorretamente. Apenas para exemplificar, ao comparar os resultados do alinhamento de *Xanthomonas campestris* pv. *campestris* contra todo o DNA, utilizando o programa TBLASTN, e todos os genes, utilizando o programa BLASTP, foi possível observar que o primeiro foi capaz de identificar diferenças em 272 grupos de genes homólogos, correspondendo a uma inconsistência de 461 genes que não foram identificados no segundo alinhamento por não terem sido anotados no banco de dados consultado. Apesar de 68% deles serem hipotéticos, diferentes tipos de genes foram identificados, incluindo diferentes tipos de proteínas do sistema de secreção que podem estar relacionadas à características patogênicas dessas bactérias.

6 Conclusões

Este trabalho teve como objetivo implementar ferramentas computacionais para identificação de genes de famílias de genes homólogos presentes nos genomas e utilizar as informações sobre os genes identificados para a realização de reconstruções filogenéticas e, com essas filogenias reconstruídas, compará-las com as filogenias obtidas em outros experimentos previamente realizados, bem como com a filogenia gerada com base na região *16S rRNA*. A partir dessas comparações, também objetivou-se identificar quais abordagens eram capazes de distinguir os diferentes genomas presentes no conjunto de dados utilizado, especialmente os genomas de organismos muito próximos evolutivamente, e que permitiriam gerar uma filogenia na qual sua topologia fosse condizente com os diferentes grupos de *Xanthomonas* analisados.

No entanto, algumas dificuldades foram encontradas ao longo do desenvolvimento do projeto. A principal foi o fato de não ser possível utilizar a filogenia da região *16S rRNA* como referência para os casos de *Xanthomonas* que fossem muito próximas evolutivamente, visto que ela não conseguiu distingui-los. Para distinguir esses genomas e para comparar os resultados dos diferentes experimentos, foi considerada como referência a filogenia gerada com base no número de genes de famílias de genes homólogos presentes em cada genoma. Entretanto, essa abordagem também apresentou limitações, visto que, apesar de diferenciar os genomas próximos e formar os três principais grupos de *Xanthomonas*, ela acabou por agrupar a *Xanthomonas albilineans* com as *Xanthomonas campestris* apesar de seu genoma ser distante dos demais de acordo com outros estudos realizados e de ser um *outlier* do conjunto de dados estudado, conforme pode ser observado nas filogenias geradas com base na região *16S rRNA* e no valor de DDH, por exemplo.

Com base nas análises e comparações das filogenias resultantes dos quatro experimentos realizados, constatou-se que não houve um consenso com relação à topologia da árvore filogenética para os genomas estudados. Algumas das abordagens permitiam distinguir os genomas mais próximos evolutivamente, mas não conseguiram posicionar a *Xanthomonas albilineans* em um ramo mais distante das demais *Xanthomonas*, enquanto que outras conseguiram considerá-la um *outlier*, mas não tiveram a capacidade de distinguir genomas muito similares. Desse modo, caso se tenha interesse em distinguir genomas de *Xanthomonas* muito próximos evolutivamente, é interessante utilizar as famílias de genes

homólogos para a reconstrução filogenética, enquanto que, nos casos em que se deseja diferenciar genomas de diferentes espécies evolutivamente distantes, a utilização da região *16S rRNA* e dos valores de DDH é a mais adequada. Há também a possibilidade de utilizar essas abordagens distintas de maneira complementar: por exemplo, utilizar a região *16S rRNA* para separar os genomas mais diferentes primeiro e depois usar as famílias de genes homólogos para diferenciar os genomas mais próximos. Além disso, pelo que foi observado no experimento realizado, as distâncias calculadas pela ferramenta GGDC podem ser utilizadas em ambas as situações, uma vez que permitiram diferenciar tanto os organismos próximos quanto os distantes evolutivamente.

As próximas seções resumem as principais contribuições obtidas neste trabalho e os possíveis trabalhos futuros para dar continuidade ao que foi realizado até o presente momento.

6.1 Principais contribuições

As principais contribuições obtidas através deste trabalho foram:

- revisão sistemática sobre a comparação de genomas completos de bactérias, publicada como um resumo ([PEREIRA; WAGNER; DIGIAMPIETRI, 2015](#));
- desenvolvimento de algoritmos para realizar a comparação automática das topologias de árvores filogenéticas;
- desenvolvimento de algoritmos para a identificação automática dos genes homólogos e geração de filogenias a partir de cada uma das famílias;
- desenvolvimento de algoritmos para seleção aleatória de genes presentes nos genomas para simular a utilização de genomas incompletos;
- experimento para geração de filogenias de genomas de *Xanthomonas* com base nos genes ortólogos presentes em todos os genomas analisados;
- experimento para reconstrução de filogenias de genomas de *Xanthomonas* com base em diferentes proporções de genes considerados para realizar essa reconstrução;
- comparação dos resultados dos experimentos realizados neste trabalho com os resultados obtidos por outros experimentos realizados.

Além da revisão sistemática realizada no contexto deste trabalho, a autora também colaborou com a revisão sistemática conduzida por outra mestranda do grupo de pesquisa

sobre a análise filogenética de genes específicos, cujos resultados obtidos também foram publicados sob a forma de um resumo (WAGNER; PEREIRA; DIGIAMPIETRI, 2015).

6.2 Trabalho futuros

Como trabalhos futuros, pode-se citar a realização de um estudo no qual se estenda o número e a variedade de genomas de *Xanthomonas*, de modo que se analise o desempenho das abordagens apresentadas neste trabalho quando são consideradas outras espécies desse gênero de bactérias e se, com esse aumento no número de genomas analisados, é possível que a *Xanthomonas albilineans* não seja agrupada com as *Xanthomonas campestris*, como ocorreu em algumas das filogenias geradas pelas abordagens apresentadas nesta dissertação. Também podem ser realizados estudos com outros procariotos para analisar o desempenho das ferramentas com genomas diferentes. Com relação aos estudos com *Xanthomonas* realizados neste trabalho, pretende-se publicar os resultados obtidos em um periódico da área.

Além disso, podem ser desenvolvidas outras abordagens para realizar a reconstrução filogenética com base nas famílias de genes homólogos identificadas pelas ferramentas apresentadas neste trabalho, por exemplo, considerando os dados das outras tabelas produzidas por essas ferramentas ou combinando as informações fornecidas por essas tabelas.

Por fim, seria interessante utilizar os resultados obtidos a partir das ferramentas desenvolvidas para a identificação de genes homólogos para se realizar uma comparação e avaliação das anotações de genes realizadas nos genomas, de modo que se tente identificar anotações realizadas incorretamente ou genes que não foram anotados.

Referências¹

- ALTENHOFF, A. M.; DESSIMOZ, C. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Computational Biology*, Public Library of Science, v. 5, n. 1, p. e1000262, 2009. Citado 2 vezes nas páginas 47 e 55.
- ALTSCHUL, S. F. et al. Basic local alignment search tool. *Journal of molecular biology*, v. 215, n. 3, p. 403–410, 1990. Citado na página 39.
- ANGIBAUD, S. et al. How pseudo-boolean programming can help genome rearrangement distance computation. In: *Proceedings of the RECOMB 2006 International Workshop on Comparative Genomics*. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg, 2006. p. 75–86. (Lecture Notes in Computer Science, v. 4205). Citado 2 vezes nas páginas 50 e 53.
- BIOLCHINI, J. et al. Systematic review in software engineering. *System Engineering and Computer Science Department COPPE/UFRJ, Technical Report ES*, v. 679, n. 05, p. 45, 2005. Citado na página 45.
- BOLSHOY, A.; VOLKOVICH, Z. Whole-genome prokaryotic clustering based on gene lengths. *Discrete Applied Mathematics*, v. 157, n. 10, p. 2370 – 2377, 2009. Citado 2 vezes nas páginas 50 e 52.
- BOOTH, K. S.; LUEKER, G. S. Testing for the consecutive ones property, interval graphs, and graph planarity using PQ-tree algorithms. *Journal of Computer and System Sciences*, v. 13, n. 3, p. 335–379, 1976. Citado na página 52.
- 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 31.
- CALLEGARI-JACQUES, S. M. *Bioestatística: Princípios e aplicações*. 1. ed. Porto Alegre: Artmed, 2008. Citado na página 95.
- CLARRIDGE, J. E. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clinical microbiology reviews*, v. 17, n. 4, p. 840–862, 2004. Citado 2 vezes nas páginas 29 e 30.
- COMIN, M.; VERZOTTO, D. Whole-genome phylogeny by virtue of unic subwords. In: *Proceedings of the 23rd International Workshop on Database and Expert Systems Applications*. United States of America: IEEE Computer Society, 2012. p. 190–194. Citado 2 vezes nas páginas 50 e 51.
- DARLING, A. E.; MAU, B.; PERNA, N. T. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE*, Public Library of Science, v. 5, n. 6, p. e11147, 06 2010. Citado 2 vezes nas páginas 50 e 53.
- DAS, S.; VISHNOI, A.; BHATTACHARYA, A. ABWGAT: anchor-based whole genome analysis tool. *Bioinformatics*, v. 25, n. 24, p. 3319–3320, 2009. Citado 2 vezes nas páginas 50 e 51.

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

DAUBIN, V.; GOUY, M.; PERRIERE, G. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome research*, v. 12, n. 7, p. 1080–1090, 2002. Citado na página 55.

DELSUC, F.; BRINKMANN, H.; PHILIPPE, H. Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*, v. 6, n. 5, p. 361–375, 2005. Citado na página 33.

DELUCA, T. F. et al. Roundup: a multi-genome repository of orthologs and evolutionary distances. *Bioinformatics*, v. 22, n. 16, p. 2044–2046, 2006. Citado na página 48.

DESSIMOZ, C. et al. OMA, a comprehensive, automated project for the identification of orthologs from complete genome data: introduction and first achievements. In: *Proceedings of the RECOMB 2005 International Workshop on Comparative Genomics*. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg, 2005. p. 61–72. (Lecture Notes in Computer Science, v. 3678). Citado na página 48.

DOMAZET-LOŠO, M.; HAUBOLD, B. Efficient estimation of pairwise distances between genomes. *Bioinformatics*, v. 25, n. 24, p. 3221–3227, 2009. Citado 2 vezes nas páginas 50 e 51.

EDGAR, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, v. 32, n. 5, p. 1792–1797, 2004. Citado na página 43.

FELSENSTEIN, J. PHYLIP - phylogeny inference package (version 3.2). *Cladistics*, v. 5, p. 164–166, 1989. Citado na página 43.

FRAZER, K. Multi-species sequence comparison: the next frontier in genome annotation. *Genome Biology*, v. 4, n. 122.1 - 122.6, p. 2677–2683, 2003. Citado na página 54.

GASCUEL, O. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular biology and evolution*, v. 14, n. 7, p. 685–695, 1997. Citado na página 56.

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 40.

GOTO, N.; KUROKAWA, K.; YASUNAGA, T. Analysis of invariant sequences in 266 complete genomes. *Gene*, v. 401, n. 1–2, p. 172 – 180, 2007. Citado 2 vezes nas páginas 50 e 51.

GUINDON, S.; GASCUEL, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology*, v. 52, n. 5, p. 696–704, 2003. Citado na página 43.

HACHIYA, T. et al. Accurate identification of orthologous segments among multiple genomes. *Bioinformatics*, v. 25, n. 7, p. 853–860, 2009. Citado 2 vezes nas páginas 49 e 50.

HAN, J.; KAMBER, M. *Data Mining: Concepts and Techniques*. 2. ed. San Francisco: Morgan Kaufmann, 2006. Citado na página 37.

HAUBEN, L. et al. Comparison of 16S ribosomal DNA sequences of all *Xanthomonas* species. *International Journal of Systematic and Evolutionary Microbiology*, v. 47, n. 2, p. 328–335, 1997. Citado 2 vezes nas páginas [29](#) e [30](#).

HOUSE, C. H.; FITZ-GIBBON, S. T. Using homolog groups to create a whole-genomic tree of free-living organisms: an update. *Journal of Molecular Evolution*, v. 54, n. 4, p. 539–547, 2002. Citado na página [59](#).

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 [30](#).

JAHN, K. Efficient computation of approximate gene clusters based on reference occurrences. In: *Proceedings of the RECOMB 2010 International Workshop on Comparative Genomics*. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg, 2010. p. 264–277. (Lecture Notes in Computer Science, v. 6398). Citado 2 vezes nas páginas [50](#) e [53](#).

JANDA, J. M.; ABBOTT, S. L. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *Journal of clinical microbiology*, v. 45, n. 9, p. 2761–2764, 2007. Citado 3 vezes nas páginas [17](#), [29](#) e [55](#).

JENSEN, L. J. et al. eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Research*, v. 36, n. suppl 1, p. D250–D254, 2008. Citado na página [47](#).

JOLLIFFE, I. *Principal Component Analysis*. 2. ed. New York: Springer, 2002. Citado na página [67](#).

KIMURA, M. *The neutral theory of molecular evolution*. New York: Cambridge University Press, 1983. Citado na página [42](#).

KITCHENHAM, B. Procedures for performing systematic reviews. *Keele, UK, Keele University*, v. 33, n. 2004, p. 1–26, 2004. Citado na página [45](#).

KORBEL, J. O. et al. SHOT: a web server for the construction of genome phylogenies. *Trends in Genetics*, v. 18, n. 3, p. 158–162, 2002. Citado na página [60](#).

KOZOBAY-AVRAHAM, L. et al. Prokaryote clustering based on DNA curvature distributions. *Discrete Applied Mathematics*, v. 157, n. 10, p. 2378 – 2387, 2009. Citado 2 vezes nas páginas [50](#) e [53](#).

KRYUKOV, K.; SAITOU, N. MISHIMA - a new method for high speed multiple alignment of nucleotide sequences of bacterial genome scale data. *BMC Bioinformatics*, v. 11, n. 1, p. 1, 2010. Citado na página [48](#).

KUBICOVA, V.; PROVAZNIK, I. Use of whole genome DNA spectrograms in bacterial classification. *Computers in Biology and Medicine*, 2014. Citado 2 vezes nas páginas [50](#) e [54](#).

KURTZ, S. et al. Versatile and open software for comparing large genomes. *Genome Biology*, v. 5, n. 2, p. 1, 2004. Citado na página [48](#).

LAING, C. R. et al. Everything at once: comparative analysis of the genomes of bacterial pathogens. *Veterinary Microbiology*, v. 153, n. 1–2, p. 13 – 26, 2011. Special Issue: VETPATH 2010. Citado na página 48.

LAM, W. W. M. et al. MAGMA: an algorithm for mining multi-level patterns in genomic data. In: *Proceedings of the 2007 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2007*. [S.l.]: IEEE Computer Society, 2007. p. 89–94. Citado 2 vezes nas páginas 49 e 50.

LAND, M. et al. Insights from 20 years of bacterial genome sequencing. *Functional & Integrative Genomics*, v. 15, n. 2, p. 141–161, 2015. Citado 3 vezes nas páginas 16, 17 e 92.

LANDAU, G. M.; PARIDA, L.; WEIMANN, O. Using PQ trees for comparative genomics. In: *Proceedings of the 16th Annual Symposium on Combinatorial Pattern Matching*. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg, 2005. p. 128–143. (Lecture Notes in Computer Science, v. 3537). Citado 2 vezes nas páginas 50 e 52.

LANG, J. M.; DARLING, A. E.; EISEN, J. A. Phylogeny of bacterial and archaeal genomes using conserved genes: supertrees and supermatrices. *PloS One*, v. 8, n. 4, p. e62510, 2013. Citado na página 57.

LARGET, B. R. et al. BUCKy: gene tree/species tree reconciliation with bayesian concordance analysis. *Bioinformatics*, v. 26, n. 22, p. 2910–2911, 2010. Citado na página 57.

LI, L.; STOECKERT, C. J.; ROOS, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research*, v. 13, n. 9, p. 2178–2189, 2003. Citado na página 48.

MEIER-KOLTHOFF, J. P. et al. Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics*, v. 14, n. 1, p. 60, 2013. Citado na página 40.

MOREIRA, L. M. (Org.). *Ciências genômicas : fundamentos e aplicações*. 1. ed. Ribeirão Preto: Sociedade Brasileira de Genética - SBG, 2015. Citado 6 vezes nas páginas 16, 24, 26, 27, 28 e 34.

MÜLLER, O. F. *Animalcula Infusoria: Fluvia Tilia Et Marina*. [S.l.: s.n.], 1786. Citado na página 6.

NEI, M.; KUMAR, S. *Molecular Evolution and Phylogenetics*. New York: Oxford University Press, 2000. Citado na página 38.

NELDER, J. A.; BAKER, R. J. Generalized linear models. *Encyclopedia of Statistical Sciences*, 1972. Citado na página 41.

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 27.

PARADIS, E.; CLAUDE, J.; STRIMMER, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, v. 20, p. 289–290, 2004. Citado na página 44.

- PEREIRA, V. M. Y.; WAGNER, P. K.; DIGIAMPIETRI, L. A. Bacterial whole genome comparison: a systematic literature review. In: *Proceedings of the X-Meeting 2015 - 11th International Conference of the AB3C + Brazilian Symposium of Bioinformatics*. São Paulo: Brazilian Association of Bioinformatics and Computational Biology (AB3C), 2015. p. 243. Citado na página 105.
- PIERETTI, I. et al. The complete genome sequence of *Xanthomonas albilineans* provides new insights into the reductive genome evolution of the xylem-limited Xanthomonadaceae. *BMC Genomics*, v. 10, n. 1, p. 616, 2009. Citado na página 101.
- QUEIROZ, A. de; GATESY, J. The supermatrix approach to systematics. *Trends in Ecology & Evolution*, v. 22, n. 1, p. 34–41, 2007. Citado na página 57.
- R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, 2016. Disponível em: <https://www.R-project.org/>. Citado na página 44.
- RODRIGUEZ-R, L. M. et al. Genomes-based phylogeny of the genus *Xanthomonas*. *BMC Microbiology*, v. 12, n. 1, p. 43, 2012. Citado na página 101.
- RUDOLPH, K. Infection of the plant by *Xanthomonas*. In: *Xanthomonas*. London: Chapman & Hall, 1993. p. 193–264. Citado na página 30.
- SCHAECHTER, M. (Org.). *Encyclopedia of Microbiology*. 3. ed. Oxford: Academic Press, 2009. Citado na página 31.
- SHARMA, V.; PATIL, P. B. Resolving the phylogenetic and taxonomic relationship of *Xanthomonas* and *Stenotrophomonas* strains using complete rpoB gene sequence. *PLoS Currents*, v. 3, p. RRN1239–RRN1239, 2010. Citado na página 101.
- SHIFMAN, A. et al. Phylo SI: a new genome-wide approach for prokaryotic phylogeny. *Nucleic Acids Research*, v. 42, n. 4, p. 2391–2404, 2014. Citado na página 61.
- SNEL, B.; BORK, P.; HUYNEN, M. A. Genome phylogeny based on gene content. *Nature Genetics*, v. 21, n. 1, p. 108–110, 1999. Citado 2 vezes nas páginas 58 e 61.
- STAMATAKIS, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, v. 22, n. 21, p. 2688–2690, 2006. Citado na página 57.
- TATUSOV, R. L.; KOONIN, E. V.; LIPMAN, D. J. A genomic perspective on protein families. *Science*, v. 278, n. 5338, p. 631–637, 1997. Citado na página 47.
- VERLI, H. (Org.). *Bioinformática: da biologia à flexibilidade molecular*. 1. ed. São Paulo: Sociedade Brasileira de Bioquímica e Biologia Molecular - SBBq, 2014. Citado 5 vezes nas páginas 16, 25, 26, 32 e 34.
- VĚTROVSKÝ, T.; BALDRIAN, P. The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PloS One*, v. 8, n. 2, p. e57923, 2013. Citado na página 29.
- WAGNER, P. K.; PEREIRA, V. M. Y.; DIGIAMPIETRI, L. A. A systematic review of phylogenetic analysis of a specific gene. In: *Proceedings of the X-Meeting 2015 - 11th International Conference of the AB3C + Brazilian Symposium of Bioinformatics*. São Paulo: Brazilian Association of Bioinformatics and Computational Biology (AB3C), 2015. p. 270. Citado na página 106.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of ‘small-world’ networks. *Nature*, v. 393, p. 440–442, 1998. Citado na página [68](#).

WAYNE, L. et al. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *International Journal of Systematic and Evolutionary Microbiology*, v. 37, n. 4, p. 463–464, 1987. Citado na página [40](#).

WOESE, C. R. Bacterial evolution. *Microbiological Reviews*, v. 51, n. 2, p. 221–271, 1987. Citado na página [29](#).

YE, Y. et al. BlastGraph: a comparative genomics tool based on BLAST and graph algorithms. *Bioinformatics*, v. 29, n. 24, p. 3222–3224, 2013. Citado na página [60](#).

ZAHA, A.; FERREIRA, H. B.; PASSAGLIA, L. M. P. *Biologia Molecular Básica*. Porto Alegre: Mercado Aberto, 2003. 424 p. Citado 2 vezes nas páginas [27](#) e [28](#).

ZHANG, Y. et al. Positive selection is the main driving force for evolution of citrus canker-causing *Xanthomonas*. *The ISME Journal*, 2015. Citado na página [30](#).