UNIVERSIDADE DE SÃO PAULO
PROGRAMA INTERUNIDADES DE PÓS–GRADUAÇÃO EM BIOINFORMÁTICA

CAIO RAFAEL DO NASCIMENTO SANTIAGO

GTACG: Um arcabouço computacional focado em genômica comparativa de bactérias de um mesmo ramo evolutivo

São Paulo
2019

CAIO RAFAEL DO NASCIMENTO SANTIAGO

GTACG: Um arcabouço computacional focado em genômica comparativa de bactérias de um mesmo ramo evolutivo

Versão corrigida

Tese apresentada à Universidade de São Paulo para obtenção do título de Doutor em Ciências pelo Programa Interunidades de Pós–graduação em Bioinformática.

Área de concentração: Bioinformática

Orientador: Prof. Dr. Luciano Antonio Digiampietri

Coorientador: Prof. Dr. Leandro Marcio Moreira

São Paulo

2019

FICHA CATALOGRÁFICA

Santiago, Caio Rafael do Nascimento

S235 GTACG: um arcabouço computacional focado em genômica comparativa de bactérias de um mesmo ramo evolutivo / Caio Rafael do Nascimento Santiago; orientador Luciano Antonio Digiampietri; coorientador Leandro Marcio Moreira. -- São Paulo, 2019.

105 f.

Tese (Doutorado) – Programa Interunidades de Pós-Graduação em Bioinformática, Universidade de São Paulo.

1. Bioinformática. 2. Genômica 3. Virulência. 4. Filogenômica. 5. Agrupamento de sequências. I. Digiampietri, Luciano Antonio, orient. II. Moreira, Leandro Marcio, coorient. III. Universidade de São Paulo. IV. Título.

CDD – 572.8

Elaborada pelo Serviço de Informação e Biblioteca Carlos Benjamin de Lyra do IME-USP, pela bibliotecária Eliana Mara Martins Ramalho CRB-8/ 4819

Tese de autoria de Caio Rafael do Nascimento Santiago, sob o título **"GTACG: Um arcabouço computacional focado em genômica comparativa de bactérias de um mesmo ramo evolutivo"**, apresentada à Universidade de São Paulo, para obtenção do título de Doutor em Ciências pelo Programa de Pós–Graduação em Bioinformática, aprovada em 25 de outubro de 2019 pela comissão julgadora constituída pelos doutores:

Prof. Dr. Luciano Antonio Digiampietri

Instituição: Universidade de São Paulo

Presidente

Profa. Dra. Aline Maria da Silva

Instituição: Universidade de São Paulo

Prof. Dr. Alessandro de Mello Varani

Instituição: Universidade Estadual Paulista

Profa. Dra. Cristina Viana Niero

Instituição: Universidade Federal de São Paulo

Dedico este trabalho a minha mãe Sandra e minha noiva Karen por estarem sempre do meu lado, mesmo nos momentos mais difíceis.

Agradecimentos

Agradeço antes de mais nada ao meu orientador, Prof. Dr. Luciano Antonio Digiampietri, por estar sempre presente de forma tão prestativa, certamente sua orientação tornou esta jornada mais tranquila.

Ao meu co-orientador, Prof. Dr. Leandro Marcio Moreira, que sem seu apoio este trabalho perderia seu embasamento biológico e, consequentemente, parte de sua razão de existir.

Às minhas colegas Renata de Almeida Barbosa Assis e Suzane de Andrade Barboza, e ao laboratório da Profa. Dra. Aline Maria da Silva pelo trabalho em fornecer informações e análises inestimáveis para a condução desta pesquisa, assim como por possibilitarem o desenvolvimento de um sistema que melhor atendesse às necessidades dos profissionais da área.

Por fim, ao CNPq pelo apoio financeiro, na forma de uma bolsa de doutorado.

Resumo

SANTIAGO, Caio Rafael do Nascimento. **GTACG: Um arcabouço computacional focado em genômica comparativa de bactérias de um mesmo ramo evolutivo**. 2019. 105 f. Tese (Doutorado em Ciências) – Programa Interunidades de Pós–Graduação em Bioinformática, Universidade de São Paulo, São Paulo, 2019.

As pesquisas no campo da genômica produzem uma grande quantidade de dados. Entretanto, o conhecimento genético acerca de certos fenótipos é limitado. Além disso, parte considerável dos genomas estudados possuem sequências codificantes (CDSs) com funções desconhecidas, representando um desafio adicional para a compreensão dos pesquisadores. Organismos provenientes de um mesmo ramo evolutivo compartilham muitas de suas CDSs, e certos fenótipos únicos a um grupo desses indivíduos podem ser resultado de um conjunto único de genes exclusivos. Neste trabalho é apresentado o arcabouço computacional GTACG, uma ferramenta com foco em uma usabilidade facilitada e destinada a pesquisas para identificação de características genéticas únicas em subgrupos de genomas de bactérias que possuem um determinado fenótipo em comum, encontrando dados que diferenciam eles dos outros organismos de forma simples. A análise do GTACG é baseada na formação de grupos de CDSs homólogas com base em alinhamentos locais. O *front-end* é simples de usar e a instalação de pacotes foi projetada para que usuários com pouco conhecimento em computação possam fazer análises complexas usando esta ferramenta. A validação dos resultados do GTACG se baseou em dois estudos de caso envolvendo um conjunto com 161 genomas da família *Xanthomonadaceae* e 45 genomas de *Streptococcus pyogenes*. No primeiro estudo de caso, buscava-se descobrir porque algumas *Xanthomonadaceae* se associam a plantas e outras não, e de fato foram encontradas 19 famílias de proteínas ortólogas exclusivas aos genomas associados a plantas (representando mais de 90% desses genomas), permitindo a identificação de proteínas potencialmente associadas com a adaptação e a virulência dessas bactérias nos tecidos das plantas. No segundo estudo, buscou-se encontrar marcadores filogenéticos para a proteína *emm* dos *Streptococcus pyogenes*, e foram encontrados 15 famílias de proteínas ortólogas que serviriam para este papel. Além disso, também foram encontrados algumas famílias combinadas que poderiam explicar parte das doenças causadas pelo *Streptococcus pyogenes* em seres humanos. Os resultados mostram o potencial de uso do GTACG para encontrar novos objetos de pesquisa para estudos moleculares de genômica comparativa de bactérias.

Palavras-chaves: Bioinformática. Genômica comparativa. Análise de pan-genomas. Fatores de virulência. Filogenômica. Agrupamento de sequências. Identificação de famílias multidomínio.

Abstract

SANTIAGO, Caio Rafael do Nascimento. **GTACG: A computational framework focused on comparative genomics of bacteria from the same evolutionary branch.** 2019. 105 p. Thesis (Doctor of Science) – Bioinformatics Graduate Program, University of São Paulo, São Paulo, 2019.

Research in the field of genomics produces a large amount of data. However, genetic knowledge about certain phenotypes is limited. Besides, a considerable part of the studied genomes has coding sequences (CDSs) with unknown functions, representing an additional challenge for researchers. Organisms from the same evolutionary branch share many of their CDSs, and certain phenotypes specific to a group of these individuals may be the result of a unique set of unique genes. In this work the GTACG computational framework is presented, a user-friendly tool to help researches to identify unique genetic characteristics in subgroups of bacterial genomes that have a common phenotype, finding data that differentiate them from other organisms in a simple way. GTACG analysis is based on the formation of homologous CDS groups based on local alignments. The front end is simple to use, and the package installation is designed to allow users with little knowledge of computer science can do complex analysis using this tool. The validation of the GTACG results was based on two case studies involving a set of 161 genomes of the Xanthomonadaceae family and 45 *Streptococcus pyogenes* genomes. In the first case study, we attempted to find out why some *Xanthomonadaceae* are associated with plants and others not, and, in fact, 19 families of orthologous proteins unique to plant-associated genomes were found (representing over 90% of these genomes), allowing the identification of proteins potentially associated with the adaptation and virulence of these bacteria in plant tissues. In the second study, we attempted to find phylogenetic markers for the protein *emm* of *Streptococcus pyogenes*, and found 15 families of orthologous proteins that would play this role. In addition, some combined families were also found that could explain some of the diseases caused by *Streptococcus pyogenes* in humans. The results show the potential use of GTACG to find new research objects for molecular studies of bacterial comparative genomics.

Keywords: Bioinformatics. Comparative genomics. Pan-genome analysis. Virulence factors. Filogenomics. Sequence clustering. Multi-domain identification.

Lista de figuras

Lista de algoritmos

Lista de quadros

Lista de tabelas

Lista de abreviaturas e siglas

CDS	*Coding Sequence* – Sequência Codificante
GTACG	*Gene Tags Assessment by Comparative Genomics*
NCBI	*National Center for Biotechnology Information*
MCL	*Markov Cluster*
COG	*Clusters of Orthologous Groups* – Clusters de Grupos Ortólogos
gb	Formato de arquivo *GenBank*
gff	Formato de arquivo *General Feature Format*
MIST	*Most Isolated SubTree*
VP	Verdadeiro Positivo
VN	Verdadeiro Negativo
FP	Falso Positivo
FN	Falso Negativo
BPGA	*Bacterial Pan Genome Analysis tool*
PGAT	*Prokaryotic Genome Analysis Tool*
PGAP	*Pan-Genomes Analysis Pipeline*
PanGP	*Pan-Genome Profile Analyze Tool*
Panseq	*Pan-genomic sequence analysis*
ITEP	*Integrated toolkit for exploration of microbial pan-genomes*
SNP	*Single Nucleotide Polymorphism* – Polimorfismo de nucleotídeo único
PRR	*Pattern-Recognition Receptors* – receptores de reconhecimento de padrões
PAMP	*Pathogen-Associated Molecular Pattern* – reconhecer padrões moleculares associados a patógenos

PTI	*Pathogen-Triggered Immunity* – ativação de gatilhos imunológicos
TBDR	*TonB-dependent receptor* – receptor TonB-dependent

Sumário

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

1 Introdução

Os estudos genéticos datam das pesquisas realizadas no início do século XX (FIETTO; LAMËGO, 2015), mas apenas na década de 70 foram apresentadas as primeiras técnicas que permitiram sequenciar trechos de DNA (SANGER; NICKLEN; COULSON, 1977). Inicialmente, o processo era bastante custoso e demorado, impossibilitando assim sua massificação. Com o passar do tempo e o avançar das tecnologias, os mecanismos disponíveis para sequenciamento genético tornaram-se mais acessíveis, rápidos e baratos, causando uma proliferação de genomas ou partes de genomas sequenciados (FIETTO; MACIEL, 2015).

O sequenciamento de genomas causou um dilúvio de dados (BELL; HEY; SZALAY, 2009). Esse cenário torna o processo de analisar todos esses dados uma tarefa consideravelmente mais custosa, visto que uma investigação manual sobre uma fração apenas de dados relevantes é progressivamente mais improvável uma vez que o surgimento desses dados é exponencial. Porém, a massificação do sequenciamento de genomas abre mais possibilidades para análises comparativas, mais especificamente de análises genômicas sobre populações (JOYCE et al., 2002; CORNEJO et al., 2013; SIMMONS et al., 2008).

A massiva quantidade de genomas sequenciados disponível nas bases de dados fornece maior grau de confiança para estudos populacionais, como a filogenia (FELSENSTEIN, 1988) e o estudo sobre fenótipos. Organismos de mesmo ramo evolutivo tendem a compartilhar determinados fenótipos, da mesma forma que compartilham determinados genes homólogos que são direta ou indiretamente responsáveis pela expressão desses fenótipos, pois a função desses genes é preservada através das gerações (HARDISON, 2003; XIA, 2013). Em geral, muitos estudos utilizam a homologia para inferir a função de proteínas desconhecidas, ou para estudar o comportamento dos organismos (CHERVITZ et al., 2011). Porém, não são abundantes os estudos que estudam populações correlatas a fim de entender os mecanismos genéticos por trás de determinados fenótipos (ILINA et al., 2013; OBOLSKI et al., 2018).

A descoberta de processos biológicos responsáveis por determinados fenótipos é uma tarefa importante para as áreas da medicina (SIMõES et al., 2015; LEE et al., 2007) e agronomia (MANSFIELD et al., 2012). Entretanto, esta não é uma tarefa fácil e, por muitas vezes, é um problema multifatorial (CASADESUS; LOW, 2006). Assim, mesmo

com os esforços da comunidade científica, a descoberta de genes específicos relacionados a determinados fenótipos ainda é um problema em aberto (CHERVITZ et al., 2011; BERGER; PENG; SINGH, 2013).

Neste contexto, a predição funcional do produto dos genes tem menos importância que a predição da região do gene em si. Essa característica é vantajosa por ser menos afetada por incongruências de anotação ou ausência de conhecimento sobre as funções. Deste modo, os problemas que ocorrem quando trabalha-se com genomas modelo de algumas espécies são evitados (LANDER et al., 2001). Assim, passa-se a estudar e entender conjuntos de genomas a partir da diversidade dessas populações (KEHDY et al., 2015). Porém, caso haja problemas na predição da sequência de aminoácidos da proteína, isso impactará negativamente nas métricas de similaridade e consequentemente vai impactar o agrupamento das proteínas em famílias homólogas.

As bactérias são organismo especialmente interessantes para essa abordagem, mais precisamente pelas suas características genômicas. Bactérias possuem genomas formados majoritariamente de sequências codificantes, diferente de organismos eucariotos que possuem a maior parte de seus pares de bases em sequências não codificantes. Por este motivo, análises que se baseiam em CDS lidam com trechos significativos do genoma quando o objeto de estudo são bactérias.

Nesta tese é apresentado o GTACG (*Gene Tags Assessment by Comparative Genomics*), um arcabouço computacional dedicado ao estudo comparativo de genomas de bactérias, mais precisamente de populações de bactérias provenientes de um mesmo ramo evolutivo (por exemplo, de uma mesma espécie, gênero ou família). Além de ferramentas comuns de arcabouços de análise de pan-genoma, o GTACG apresenta ferramentas específicas para o estudo de fenótipos presentes em uma parcela da população estudada.

1.1 Arcabouços de genômica comparativa

De maneira geral, arcabouços computacionais e *pipelines* (execuções em sequência) possuem três fases bem definidas: pré-processamento, processamento e visualização de resultados.

1.1.1 Pré-Processamento

Os arcabouços que se baseiam em relações de homologia precisam realizar algum tipo de pré-processamento sobre as CDS. Uma das etapas mais importantes desse processo é a (re)anotação automática das CDS. Isto é necessário pois os genomas disponíveis em bases de dados públicas, como o *National Center for Biotechnology Information* (NCBI), possuem dados submetidos por indivíduos diferentes, os quais utilizaram metodologias diferentes de anotação. Isto poderia causar problemas durante as etapas posteriores de processamento, análise e visualização dos dados (SETUBAL; WATTAM; ALMEIDA, 2018). Apesar de ser uma etapa bastante importante, para a maioria dos arcabouços ela não é obrigatória, estando a cargo do usuário optar por realizá-la ou não.

A etapa de pré-processamento seguinte é a identificação de famílias de genes homólogos. Nesta etapa as sequências de proteínas são agrupadas. Algoritmos de agrupamento (ou clusterização) de sequências costumam se basear em métricas de similaridade entre as sequências, como identidade, percentual de alinhamento ou até mesmo *k-mers* (COMIN; VERZOTTO, 2012). Por esse motivo, alguns desses métodos utilizam alinhamentos locais entre todas as CDS do conjunto contra todas as CDS. Ao final deste processo espera-se que subconjuntos de CDS sejam formados por elementos com alta similaridade entre si.

Existem diversas técnicas bem conhecidas na literatura deste tema (HAN; KAMBER; PEI, 2011), cada uma com propósitos e limitações diferentes. Entre as diversas classes de algoritmos de agrupamento, os métodos mais empregados para agrupar famílias de genes são baseados em vizinhança ou em grafos.

O princípio de ambos os métodos é bastante similar, baseiam-se em abordagens *bottom-up* ou *top-down*. Os métodos bottom-up começam com agrupamentos (ou *clusters*) unitários ou pequenos, e seguindo critérios de densidade aglutinam os menores para formar novos agrupamentos. A abordagem bottom-up de ligação simples ou única (*single-linkage*) é uma das mais rápidas, basta uma única relação entre sequências de agrupamentos diferentes para aglutinar dois agrupamentos. GeneRage (ENRIGHT; OUZOUNIS, 2000) é um exemplo de abordagem que utiliza *single-linkage* em que basta haver similaridade (dentro de limiares definidos) entre dois genes para juntar dois agrupamentos. Por outro lado, a técnica *complete-linkage* exige que exista alta similaridade entre todos os genes de

dois agrupamentos para aglutiná-los. Essa abordagem é mais completa e evita agrupamentos com similaridade média baixa, mas tem desvantagem de ser um pouco mais lenta. Medidas estatísticas sobre a similaridade dos conjuntos costumam ser utilizadas nas abordagens que utilizam *complete-linkage*, como os métodos apresentados por Sasson, Linial e Linial (2002) que usam média aritmética, geométrica e harmônica, ou como o método apresentado por Abascal e Valencia (2002) que utiliza entropia. O processo de aglutinar clusters se repete até o momento que a junção de dois clusters diminua a qualidade da clusterização.

O algoritmo MCL (*Markov Cluster Algorithm*) (DONGEN, 2000) se baseia em uma abordagem de vizinhança, assim como o *complete-linkage*, porém ele considera a vizinhança de forma não determinística. O algoritmo utiliza Modelos de Estados Ocultos de Markov (em inglês *Hidden Markov Models*) para realizar passeios aleatórias sobre o grafo no qual os nós correspondem a CDS e arestas correspondem a relações de similaridade entre CDS. O MCL foi proposto inicialmente para agrupar dados de origem biológica, e com o desenvolvimento do TribeMCL (ENRIGHT; DONGEN; OUZOUNIS, 2002) ele foi adaptado para o agrupamento de sequências. Atualmente, o MCL é um dos algoritmos de agrupamento de sequências mais utilizados, se mostrando bastante robusto e veloz e pouco afetado por pequenas mudanças de topologia (BROHÉE; HELDEN, 2006), além de também servir de base para outros serviços como o OrthoMCL (LI, 2003), Roary (PAGE et al., 2015), GET_HOMOLOGUES (CONTRERAS-MOREIRA; VINUESA, 2013) e PanX (DING; BAUMDICKER; NEHER, 2018).

Por mais que utilizados em outros domínios de conhecimento, os métodos baseados em agrupamento *top-down*, em que os grupos são inicialmente muito grandes para em seguida serem divididos, não costumam ser utilizados para o agrupamento de sequências. Isso se deve ao fato de que até mesmo um experimento relativamente simples de agrupamento de sequências possuírem dezenas ou centenas de milhares de genes, podendo ter dezenas de milhões de alinhamentos. Neste contexto, seria bastante custoso o uso desse tipo de abordagem que parte de grupos grandes para em seguida subdividi-los.

1.1.2 Processamento

Alguns arcabouços e *pipelines* realizam etapas adicionais ao agrupamento para separar as famílias de genes encontradas em conjuntos ortólogos (em inglês *Clusters of*

Orthologous Groups – COG). Estas etapas podem utilizar um método próprio que têm a ortologia como cerne, como o InParanoid (O'BRIEN; REMM; SONNHAMMER, 2005) ou como o trabalho de Fa Zhang et al. (2005), os quais são aplicados mais comumente a sequências de eucariotos.

Parte dos arcabouços utiliza métodos já prontos de agrupamento de sequências para produzir COGs, com especial destaque para o MCL que é utilizado no Roary (PAGE et al., 2015), OrthoMCL (LI, 2003), GET_HOMOLOGES (CONTRERAS-MOREIRA; VINUESA, 2013), PanX (DING; BAUMDICKER; NEHER, 2018). Uma abordagem utilizada pelo Roary e o OrthoMCL (entre outros) é baseada em similaridade: dada uma família de genes identifica-se o número máximo de cópias em um genoma. Essas n cópias são separadas em n COGs diferentes e então os demais genes agrupados na família tentam ser encaixados nesses n COGs da melhor forma possível. Uma abordagem alternativa é utilizar como base a filogenia da família em questão para "cortar" ramos com base em critérios de exclusão de parálogos, como feito pelo PanX ou no trabalho de Setubal, Stoye e Stadler (2018).

Uma etapa subsequente comum em alguns arcabouços computacionais é produzir alinhamentos múltiplos e filogenias para cada uma das famílias encontradas. Essa etapa consome um tempo considerável e, especialmente por esse motivo, não é empregada em todos os arcabouços. Arcabouços que se propõem a ser mais rápidos não costumam produzir alinhamentos ou filogenias. Também por esse motivo que, quando utilizado, opta-se por programas mais rápidos, mesmo que os resultados produzidos não sejam os de maior acurácia, como o FastTree (PRICE; DEHAL; ARKIN, 2010).

Nesta etapa também são produzidos resultados comuns a análises de pan-genoma, como a quantidade de famílias do core e do pan-genoma, assim como o gráfico com o perfil de crescimento do pan-genoma (TETTELIN et al., 2008). Outro resultado comum é a lista de famílias de genes, indicando a presença (marcada como 1) ou ausência (marcada como 0) de genes representantes de cada um dos genomas, e/ou uma lista com genes exclusivos/acessórios.

A filogenia também é uma linha de pesquisa comum a diversos arcabouços computacionais de genômica comparativa. Existe uma ampla gama de técnicas de inferência de filogenia, sendo que algumas são pouco utilizadas em alguns problemas de genômica comparativa (DELSUC; BRINKMANN; PHILIPPE, 2005) devido à grande quantidade de informação do problema. Este é o caso dos métodos que produzem um alinhamento

múltiplo de todos os cromossomos. O alinhamento múltiplo é muito custoso e é mais recomendado para sequências menores como genes. Por esse motivo é mais comum utilizar apenas o alinhamento de trechos específicos, como marcadores filogenéticos (HAUBEN et al., 1997). Outra opção é concatenar os genes do core-genoma e alinhá-los. Ainda com base nos cromossomos, os métodos que fazem comparações livres de alinhamentos costumam ser consideravelmente mais rápidos (LEIMEISTER et al., 2014), desta forma é possível fazer uma matriz de distância entre os genomas de interesse e, com base nela, produzir a filogenia.

Considerando as limitações de recurso associadas à utilização de sequências muito grandes, a abordagem baseada em famílias de genes se torna mais leve e veloz. Esta abordagem permite, por exemplo, a utilização da lista de presença/ausência de genes em cada uma das famílias para o cálculo de uma matriz de distância. Outra possibilidade é a produção de uma *supertree* (CREEVEY; MCINERNEY, 2009) que sumarize as relações filogenéticas de todas as filogenias produzidas para cada uma das famílias.

1.1.3 Visualização dos resultados

Existem três formas predominantes de se apresentar os resultados ao usuário. A primeira e mais simples é a apresentação de resultados na forma de arquivos de texto. Ocasionalmente o usuário precisará fazer uso de algum outro software para a visualização dos dados, e por mais que isso não seja o mais cômodo, essa abordagem é importante para usuários com maior conhecimento em computação (ou em ferramentas de bioinformática) e com capacidade de criar seus próprios *pipelines*.

Os resultados também podem ser apresentados por meio de aplicações para computador ou *web sites*. As aplicações para computador possuem maior disponibilidade de recursos já desenvolvidos, enquanto os web sites não estão limitados à configuração do usuário, isto é, não são dependentes de sistemas operacionais ou bibliotecas instaladas, além de serem uma forma mais dinâmica de compartilhar resultados entre pesquisadores.

1.2 Motivação

Por mais que existam diversas ferramentas e arcabouços computacionais para auxiliar na análise comparativa de genomas, parte considerável deles se limita a descrições estatísticas sobre o pan-genoma. Esses dados estatísticos como o tamanho do core-genoma ou a quantidade de genes exclusivos de um determinado organismo dão uma visão geral importante sobre uma população, porém há outras informações que poderiam ser exploradas. O pan-genoma abrange uma grande quantidade de informação, e por mais que o core-genoma seja relevante à categorização de uma espécie ou de qualquer outro ramo evolutivo, existe ainda uma quantidade maior de informação presente nos genes acessórios que não costumam ser estudados na maioria dos arcabouços. Desta forma, pesquisas considerando genes acessórios podem ser importantes para a categorização de um subconjunto dos organismos estudados.

Por esse motivo é apresentado nesta tese um novo arcabouço computacional para enriquecer o conhecimento sobre os organismos utilizando informações de todos os genes (incluindo os acessórios). A partir de uma anotação sobre dados fenotípicos (realizada manualmente por especialistas de domínio, experimentos, etc), considerando que existe um subgrupo de genomas que apresenta um dado fenótipo e o restante que não apresenta, pode-se investigar as famílias de genes a fim de encontrar aquelas que melhor refletem os padrões que foram estabelecidos na anotação fenotípica.

1.3 Objetivos

O objetivo principal deste trabalho foi especificar e desenvolver um arcabouço computacional para análise genômica de bactérias, tendo como base as relações de homologia. Para isso foi necessário, além de fornecer as ferramentas de análise, também prover uma estrutura de dados para o armazenamento das sequências de forma eficiente, assim como do relacionamento entre elas.

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

- Aperfeiçoar um método de agrupamento/clusterização de sequências tendo em vista as características próprias do problema, isto é, CDS obtidas de genomas completos de bactérias de um mesmo ramo evolutivo;

- Fornecer as ferramentas necessárias para a construção de filogenias utilizando as informações referentes à filogenia específica dos grupos de sequências;
- Prover análises que possam determinar diferenças entre um e outro genoma, ou grupos de genomas;
- Desenvolver um método para armazenar os genomas em grupos de interesse e desenvolver meios para analisar os genomas no contexto desses grupos;
- Apresentar os resultados gerados na forma de um web site estático (todos os processamentos principais realizados previamente), para que dessa forma sejam facilmente publicados e compartilhados;
- Apresentar os resultados obtidos na forma de estudos de casos e verificar a validade ou importância dos resultados encontrados em relação a resultados já analisados na literatura correlata.

1.4 Organização do texto

O restante deste texto está organizado de modo que este capítulo apresenta a introdução ao tema abordado e conceitos básicos, assim como os objetivos estabelecidos. O capítulo 2 descreve os materiais e métodos utilizados no desenvolvimento deste projeto. Já o capítulo 3 descreve quais foram os resultados obtidos, detalhando o funcionamento do arcabouço desenvolvido, além de uma descrição dos estudos de casos realizados. O capítulo 4 apresenta a análise dos resultados obtidos. Por fim, o capítulo 5 apresenta as conclusões do projeto e uma breve discussão sobre possíveis desdobramentos futuros.

2 Materiais e métodos

O ambiente como um todo do GTACG pode ser dividido entre *back-end* e *front-end*. A divisão se faz necessária, pois no back-end estão as ferramentas e algoritmos destinados a preparação dos dados genômicos fornecidos pelo usuário e é necessário um conhecimento básico em computação para executar as etapas de seu *pipeline*. Por outro lado, o front-end é destinado à visualização dos resultados e não exige conhecimentos específicos em computação. O back-end foi desenvolvido em Java, pois trata-se de uma linguagem de programação portável entre diferentes sistemas operacionais e de fácil execução. Já o front-end foi desenvolvido em HTML e JavaScript.

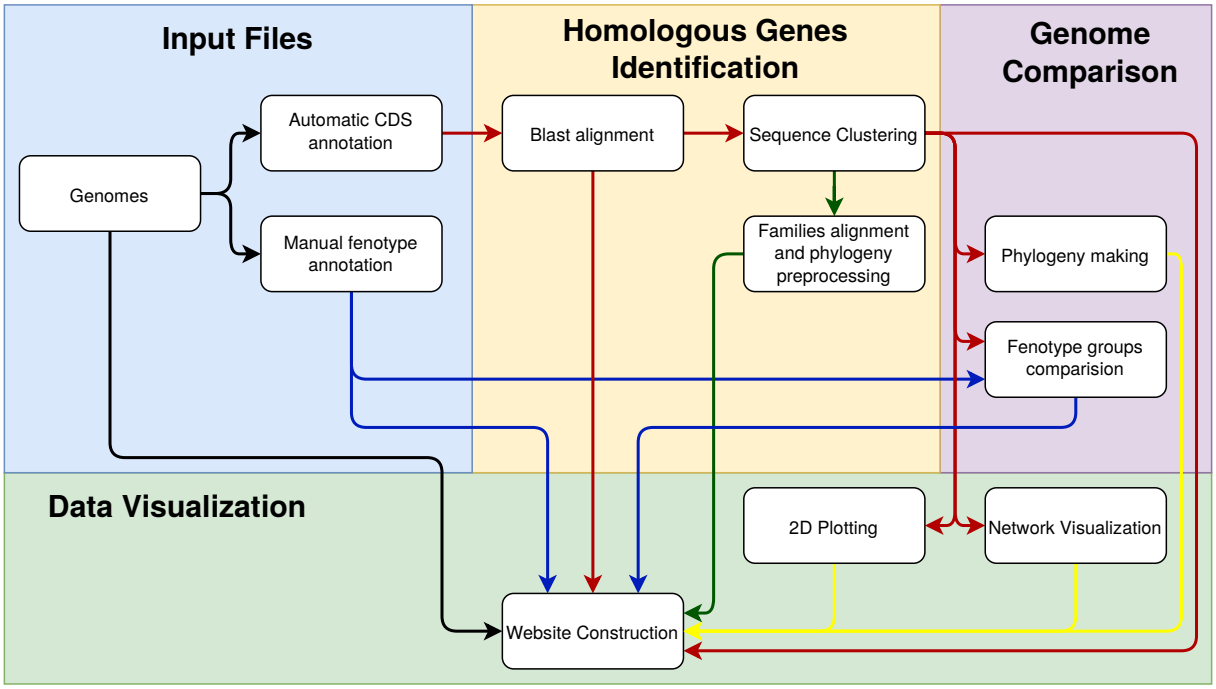
Entre os arquivos de entrada necessários estão: (1) arquivos com o sequenciamento completo dos genomas (no formato *FASTA*), (2) arquivos contendo a anotação dos genomas (no formato de texto simples para, por exemplo, identificar características fenotípicas) e (3) arquivos contendo as anotações das CDS (preferivelmente feita de forma automática nos formatos FASTA, *gb*, *gbf* ou *gff*). A execução do GTACG segue os passos como descritos na figura 1, e as etapas se focam em três principais pilares: (1) identificação de genes homólogos, (2) comparação de genomas completos, e (3) visualização de dados. A reanotação automática das CDS não é parte integral do pipeline, porém ela é importante para evitar inconsistências metodológicas. Cabe ao usuário realizar essa tarefa de pré-processamento com a ferramenta ou serviço de sua escolha. Nas seções seguintes os três pilares serão discutidos em seus detalhes.

2.1 Identificação de genes homólogos

A identificação de famílias de genes homólogos é uma tarefa bastante importante em uma análise de genômica comparativa. Os resultados obtidos nessa etapa impactam a maioria das conclusões subsequentes. Por este motivo foi proposto um algoritmo de agrupamento de sequências com um enfoque em sequências provenientes de genomas distribuídos em um mesmo ramo evolutivo.

O princípio que norteou o desenvolvimento deste algoritmo foi a transitividade das relações entre sequências homólogas. Como exemplificado pela figura 2, a homologia por definição é uma relação transitiva (SASSON; LINIAL; LINIAL, 2002), isto é, caso

Figura 1 – Etapas envolvidas na execução do *pipeline* do GTACG organizadas em três pilares de processamento: identificação de genes homólogos, comparação de genomas e visualização de resultados. Para facilitar a visualização das relações entre os dados, cada um deles foi colorido de acordo com o seguinte esquema: em preto estão dados gerais sobre os genomas; em azul estão os dados referentes a grupos de genomas; em vermelho estão os dados de sequências; e em amarelo estão resultados gráficos para visualização.

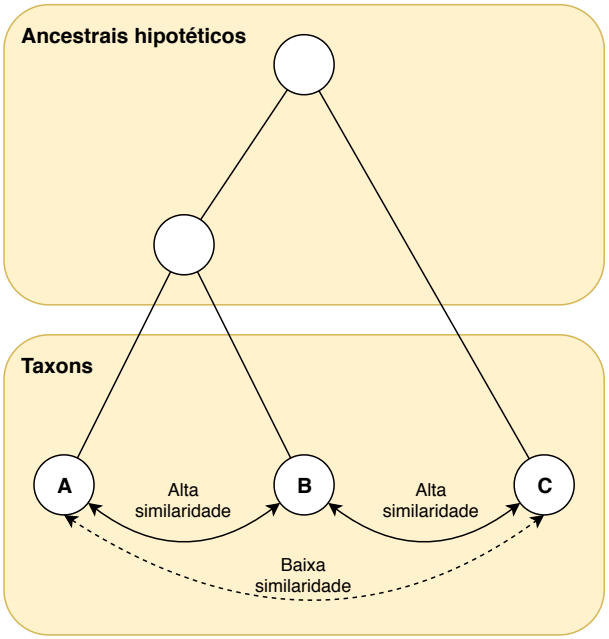


Fonte: Santiago et al. (2019)

duas sequências quaisquer A e B sejam definidas como homólogas, isto é, que haja um ancestral hipotético comum a elas, e ficando provada a homologia entre B e C, por definição deve haver um ancestral hipotético comum a A, B e C. Entretanto, o agrupamento de sequências não se dá por uma categorização estrita de homologia, mas sim por relações de similaridade, por serem mais facilmente estabelecidas por meios matemáticos. Métricas de similaridade não são transitivas, assim, no mesmo cenário do exemplo anterior, caso A e B tenham alta similaridade, e o mesmo aconteça com B e C, isso não implica em uma alta similaridade entre A e C.

A maioria dos algoritmos de agrupamento de sequências é baseada em medidas de similaridade. Assim, algoritmos baseados em grafos ou grupos vizinhos, que usam esse tipo de medida, podem produzir resultados com problemas derivados da transitividade. Ao final do agrupamento, sequências que estão em uma mesma componente conexa ou grupo de vizinhos são consideradas homólogas, mas esses grupos podem ter sido formados por

Figura 2 – Exemplo da relação transitiva de homologia e seu paralelo não transitivo em relação à similaridade. Havendo uma relação de homologia entre A e B, e entre B e C e sendo esta propriedade transitiva, haverá também homologia entre A e C. Porém, utilizando como relação a similaridade entre sequências, esse comportamento não é necessariamente verdadeiro. Isto é, o fato das sequências A e B serem similares (dentro de certos limiares) entre si e B e C também serem similares, não garante que as sequências A e C também serão similares (considerando os mesmos critérios/limiares).



Fonte: Caio Santiago, 2019

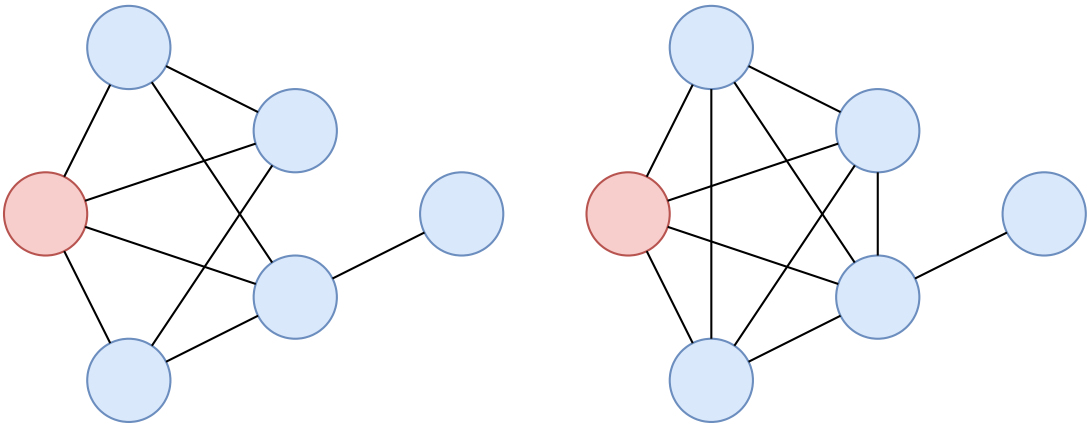
relações de alta similaridade que ocasionaram a criação de famílias que possuem muitas relações de baixa similaridade entre suas sequências. Basta uma única relação de alta similaridade para agrupar duas famílias homólogas não correlatas.

O algoritmo proposto para o agrupamento de sequências tem como objetivo minimizar conexões baseadas puramente em transitividade, diminuindo assim os casos em que famílias de genes homólogos contenham muitas sequências com baixa similaridade entre si. Entende-se, portanto, que famílias idealmente aceitas como homólogas tenham todas as sequências com alta similaridade com as restantes, do ponto de vista de teoria de grafos, idealmente, todos os vértices estariam ligados a todos os outros vértices (sendo denominado dessa forma como grafo completo).

O algoritmo desenvolvido utiliza o coeficiente de agrupamento (*clustering coefficient*) (WATTS; STROGATZ, 1998) para criar componentes conexas mais densas e mais fortemente baseadas em relações de alta similaridade. O coeficiente de agrupamento (ou coeficiente de clusterização) é uma métrica topológica para grafos em que, para cada con-

junto de três vértices conectados, é calculada a probabilidade desses três vértices formarem um clique de tamanho três, isto é, os três vértices estarem todos diretamente conectados entre si, como exemplificado pela figura 3. Quanto mais densas forem as componentes do grafo, mais próximo de 1 tende a ser o coeficiente; da mesma forma que quanto mais esparsas forem as componentes do grafo, mais próximo de 0 tende a ser o coeficiente. Uma característica importante deste coeficiente é que ele não é negativamente impactado pela quantidade de componentes isoladas. Se todas as componentes do grafo formem grafos completos, então o coeficiente de agrupamento para todo o grafo será 1.

Figura 3 – Exemplificação visual do coeficiente de agrupamento. Analisando apenas os vértices do grafo que estão destacados em vermelho, o vértice da componente a esquerda possui coeficiente de agrupamento igual a 0,66, pois dos seis possíveis cliques de tamanho três que envolvem o vértice em destaque, apenas quatro desses cliques se concretizam. Já o vértice em destaque da componente a direita possui coeficiente de agrupamento igual a 1, pois todos os cliques de tamanho três envolvendo o vértice em destaque se concretizam.



Fonte: Caio Santiago, 2019

O algoritmo desenvolvido se baseia em métricas de similaridade sobre alinhamentos locais, produzidas por ferramentas como o BLASTP (CAMACHO et al., 2009) ou MMSeqs2 (STEINEGGER; SÖDING, 2017). Logo, o primeiro passo é produzir os alinhamentos locais de todas as sequências contra todas. Esta é uma etapa comum a algoritmos de agrupamento de sequências (ABASCAL; VALENCIA, 2002; ENRIGHT; DONGEN; OUZOUNIS, 2002; SASSON; LINIAL; LINIAL, 2002), sendo inclusive comum limitar os resultados a apenas alinhamentos com e-value menores que 10^{-5} ou 10^{-10}, pois alinhamentos com valores maiores dificilmente indicariam uma relação de homologia a ser considerada. Nesta tese, o limiar máximo definido para o e-value foi 10^{-10}. Além disso, foi verificado empiricamente que os resultados melhoram consideravelmente quando uma porcentagem mínima sobre o tamanho do alinhamento (considerando o tamanho da maior

sequência) era definida. Outros limiares poderiam ser definidos, conforme a escolha de quem utiliza o algoritmo, como por exemplo, a porcentagem mínima de identidade ou o número máximo de lacunas (*gaps*).

Os resultados que satisfazem as condições definidas são transformados em um grafo, no qual os vértices representam as sequências e as arestas representam os alinhamentos. As componentes conexas do grafo são consideradas como grupos homólogos, entretanto nesta etapa ainda é preciso retirar as arestas que, provavelmente, não representam boas relações de homologia.

A etapa seguinte é a de retirada de arestas para que o grafo assuma uma topologia mais próxima da desejada (distribuição mais homogênea e componentes mais densas). Pelo fato de todos os genomas estarem em um mesmo ramo evolutivo, espera-se que o grafo resultante seja mais homogêneo e denso e, em um caso ideal, que as componentes do grafo sejam todas completas (nas quais todos os nós estão ligados entre si), pois uma CDS homóloga estará (com variações pequenas) em parte significativa dos genomas, enquanto que CDS não homólogas não encontrarão correspondência nos outros genomas. Para isso é escolhido o limiar de corte para o e-value entre 10^{-10} e 10^{-180} para que sejam excluídas todas as arestas que representem alinhamentos com e-value maiores e, por isso, são mais prováveis de não representarem relações homólogas. Esse limiar de corte é definido como a valor que maximize a média do coeficiente de agrupamento de todos os vértices (coeficiente de agrupamento médio local) dentro do intervalo, porém a busca dentro do intervalo contínuo é muito custosa computacionalmente e, por isso, a busca é realizada dentro de intervalo discreto de n valores, no caso, foram utilizados 171 valores: $[10^{-10}, 10^{-11}, 10^{-12}, \ldots, 10^{-180}]$. O custo computacional para se encontrar esse valor, utilizando um algoritmo simples, seria de $\mathrm{O}(n * (|V|^2) - n * |V|)$, sendo n o número de limiares testados e $|V|$ o número de vértices (sequências do grafo). Porém, com o uso de programação dinâmica o custo é reduzido para $\mathrm{O}(|V|^2 - |V| + n)$.

Encontrado o limiar de corte sobre o e-value que maximiza o coeficiente de agrupamento, remove-se todas as arestas com alinhamentos cujos e-value sejam maiores do que esse valor. Em seguida, as componentes que estão completas (em que todas as sequências estão completamente ligadas entre si) são separadas, e o processo se repete até que nenhum novo limiar de corte melhore o coeficiente de agrupamento do grafo como um todo. Ao final do processo resta apenas uma lista de diferentes limiares de corte que se tornam progressivamente mais restritivos e indicam como as arestas devem ser retiradas para se

chegar ao grafo de homologias. O pseudocódigo que realiza estas operações pode ser visto no algoritmo 1.

Como o algoritmo de agrupamento de sequências preserva as relações entre os vértices do grafo, isso abre diversas possibilidades de análises topológicas. Uma delas é a identificação de possíveis domínios e *motifis*, que são muito relevantes para estudos genéticos (VOGEL et al., 2004).

Sequências multidomínio são amplamente conhecidas por serem um problema para algoritmos de agrupamento (VOGEL et al., 2004). Elas podem ser responsáveis pelo agrupamento de sequências de maneira errônea pelo fato de serem compostas de trechos de sequências provenientes de mais de um grupo homólogo. Essa situação é bastante problemática porque leva o algoritmo a produzir grupos com sequências não homólogas. Isto pode ser observado no exemplo apresentado pela figura 4, em que uma única sequência apresenta alta similaridade com dois grupos bem definidos de sequência homólogas, fazendo com que os dois grupos sejam entendidos como uma mesma família de genes. Do ponto de vista de teoria dos grafos e análise topológica, esse tipo de sequência produz vértices com coeficientes de agrupamento menores que os de seus vizinhos (assim como a sequência apresentada na figura 4, ilustrada no grafo da figura 5).

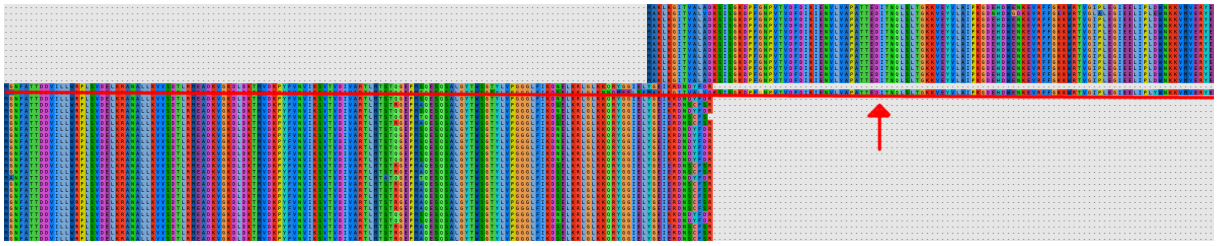
Logo, um primeiro passo para a identificação de domínios é identificar todos os vértices que possuem um coeficiente de agrupamento que seja menor que a média do coeficiente de seus vizinhos. Esses vértices são marcados como possivelmente multidomínio, por outro lado os demais são marcados como de domínio único.

Algoritmo 1 Algoritmo de agrupamento de sequências baseado no coeficiente de agrupamento médio local

```
1:  procedure CLUSTERIZACAO(Grafo g, Inteiro início, Inteiro fim)
2:      listaDeCortes = ∅
3:      while true do
4:          corte = max_{i=início}^{fim} (coeficienteClusterizacao(g, i))
5:          if corte = início then return listaDeCorte
6:          início = corte
7:          listaDeCortes ← corte
8:          aplicarCorte(g, corte)
9:          Grafo novo
10:         for sub in componentes(g) do
11:             if |vértices(sub)| > 2 & coeficienteClusterizacao(sub) < 1 then
12:                 novo ← sub
```

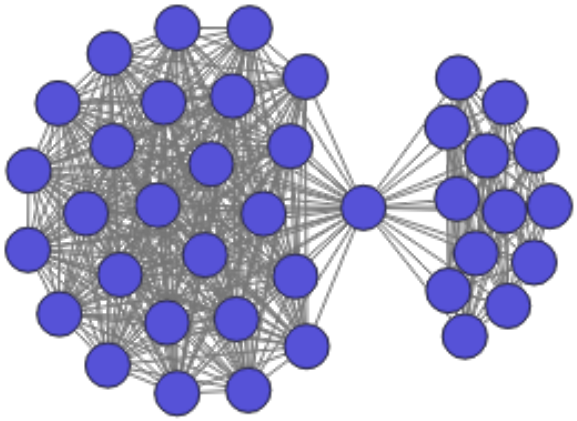
Fonte: Santiago, Pereira e Digiampietri (2018)

Figura 4 – Alinhamento correspondente a uma família de genes em que uma única sequência apresenta alta similaridade com dois grupos bem definidos de proteínas homólogas. A sequência está destacada com uma seta vermelha e sublinhada em vermelho.



Fonte: Caio Santiago, 2019

Figura 5 – Grafo das relações de similaridade destacando um caso específico de uma proteína multidomínio. Neste cenário, uma única sequência tem alta similaridade com todas as demais, mas possui baixo coeficiente de agrupamento. Neste caso nós que se destaca ao centro por estar isolado dos demais tem coeficiente de clusterização baixo, enquanto que os outros possuem o coeficiente de clusterização 1 (valor máximo desta métrica).



Fonte: Caio Santiago, 2019

Em seguida o grafo passa por um processo de simplificação. Vértices conexos de domínio único (que não dependem de relações com os multidomínio para se conectarem) são convertidos em um único vértice (um representante simbólico do grupo), preservando suas arestas e os valores de seus alinhamentos locais entre os vértices do grupo para os que estão fora. O mesmo é feito com as sequências multidomínio, as que são conexas entre si são convertidas em um único vértice preservando suas arestas e alinhamentos.

A próxima etapa transforma as arestas para que se tornem direcionais. Para isso, todas as arestas são verificadas e se todos os alinhamentos locais de um vértice a outro forem maiores que dois parâmetros pré-definidos, então a aresta será direcionada (partindo da menor sequência em direção à maior). Os dois parâmetros são baseados em diferenças

no comprimento do alinhamento entre duas sequências, o primeiro é o valor absoluto da diferença e o segundo é a diferença dividida pelo tamanho do alinhamento. Em uma análise empírica, os valores 100 e 0,3 se mostraram adequados para resolver o problema de maneira satisfatória, alcançando os resultados apresentados e discutidos nos próximos capítulos.

Grafos direcionados implicam que nem sempre é possível acessar todos os vértices a partir de um determinado vértice de início (em uma componente conexa). Portanto, a partir de um determinado vértice de início há um subconjunto do conjunto total de vértices da componente conexa que são acessíveis (incluso o próprio vértice de início). Os diferentes conjuntos que são acessíveis a partir de diferentes vértices de início serão considerados diferentes domínios (conjuntos repetidos são desprezados). Caso a componente conexa seja também fortemente conexa, então neste caso todos os conjuntos de vértices acessíveis são iguais ao conjunto de todos os vértices da componente conexa independente do início estabelecido e, assim, ela não é considerada uma componente multidomínio.

Em paralelo à identificação de famílias multidomínio, é feita uma análise simples de famílias ortólogas. Dada a árvore filogenética das famílias, caso algum ramo da árvore seja maior que um determinado limiar o ramo é excluído, transformando assim uma família homóloga em duas ou mais famílias ortólogas (DING; BAUMDICKER; NEHER, 2018).

Por fim, para cada uma das famílias, nos três níveis de profundidade (homologia, ortologia, e domínios), são produzidos alinhamentos múltiplos e a respectiva filogenia, utilizando as ferramentas Clustal Omega (SIEVERS et al., 2011) ou Muscle (EDGAR, 2004) e FastTree (PRICE; DEHAL; ARKIN, 2010) ou PhyML (GUINDON et al., 2010).

2.2 Comparação de genomas completos

Três diferentes abordagens foram utilizadas para inferir a filogenia dos genomas completos, tomando como base as famílias de genes calculadas na etapa anterior. A primeira abordagem utilizada considera a presença e ausência de CDS de cada genoma nas famílias de genes para gerar um vetor binário de características. Caso o genoma tenha uma de suas CDS em determinada família assume-se o valor 1, e 0 caso contrário. A junção de todos os vetores de características (respectivos a cada um dos genomas) é passada ao Phylip (FELSENSTEIN, 2005) para inferir a filogenia dos genomas. A segunda abordagem

usa uma matriz de distância para inferir a filogenia por meio do algoritmo Neighbor-Joining presente no Phylip. Esta matriz é construída através da distância Euclidiana entre os vetores de características binários. Por fim, a terceira abordagem se baseia na sumarização de relações filogenéticas entre um conjunto de filogenias. O conjunto de filogenias em questão é referente às filogenias de cada uma das famílias (geradas a partir dos alinhamentos das sequências). As filogenias provenientes do core-genoma são utilizadas para calcular o consenso, já as filogenias do pan-genoma são utilizadas para a *supertree* (CREEVEY; MCINERNEY, 2005).

Com relação a busca de características relacionadas à anotação prévia dos genomas (por exemplo, informações fenotípicas), a abordagem utilizada consiste em encontrar características que são comuns a determinado grupo de genomas (que compartilhem uma mesma anotação ou rótulo) e ao mesmo tempo incomum aos genomas de fora do grupo. Com base neste princípio foram produzidos dados de cada uma das famílias em possíveis categorias. A primeira dessas categorias é a conformação das famílias, definida por famílias (individualmente ou em combinação) únicas ou majoritárias a um grupo específico de genomas. Nesta categoria são apresentadas métricas para indicar quantas CDS ou genomas presentes na família pertencem aos genomas dos grupos de interesse. A informação é disponibilizada na forma absoluta e percentual, indicando deste modo o quão representativo este grupo é para a família em questão. A segunda categoria apresentada é referente aos alinhamentos das famílias, identificando de forma relativa quantas bases são mais correlatas a determinado grupo, e para expressar isso de forma numérica foi criada uma métrica de dissimilaridade. A última categoria é sobre as filogenias inferidas a partir dos alinhamentos múltiplos das sequências de cada família, com o objetivo de determinar o quão bem estão separados os genomas de determinado grupo em relação aos outros. Para isso foi criada a métrica *Most Isolated SubTree* (MIST) que mostra de forma numérica qual o tamanho da maior sub-árvore formada apenas por sequências do grupo em estudo.

2.3 Visualização de dados

Da mesma forma que a comparação de genomas, a visualização dos dados é bastante dependente da conformação das famílias. Como o algoritmo para a identificação das famílias de genes utiliza uma abordagem baseada em grafos (e mais particularmente o

resultado preserva as arestas originais) é possível a apresentação do pan-genoma como uma rede gênica. Nesta rede as famílias são representadas como componentes conexas, provendo uma noção mais clara da distribuição do pan-genoma. Um algoritmo do tipo *force-directed* (KOBOUROV, 2012) foi utilizado para aproximar ou separar as sequências com base nas suas arestas, permitindo assim uma visualização bidimensional desses dados sem que haja muita sobreposição de vértices.

Um mapeamento bidimensional dos genomas também foi realizado utilizando a mesma matriz de distância construída a partir dos vetores binários de características utilizados na inferência filogenética. Com base em um algoritmo de *Multidimensional Scaling* (BORG; GROENEN, 2005), a matriz de distância é aproximada para o plano bidimensional, preservando proporcionalmente as distâncias presentes na matriz, resultando em uma visão geral da distância entre os genomas analisados.

Nesta etapa, todos os dados calculados até o momento são consolidados na forma de um web site estático, isto é, não há necessidade de configurações complexas de servidores ou de sistemas operacionais por parte do usuário para poder usufruir da maioria das funcionalidades do sistema. Isso se deve ao fato dos dados já terem sido pré-processados. O web site não necessita de um sistema de banco de dados porque os dados são gerenciados por códigos escritos em *JavaScript* e armazenados como conjuntos de arquivos.

O formato de web site foi escolhido com base em algumas qualidades, entre elas está a facilidade de compartilhar os resultados com outros colaboradores ou publicamente com a comunidade científica. O ambiente como um todo pode ser facilmente modificado e estendido, e por ser escrito em HTML/JavaScript (um padrão já consolidado da internet) não é necessário conhecimento específico em outros arcabouços computacionais para implementar novas funcionalidades. Como os dados são exportados para arquivos no formato JavaScript, é relativamente fácil copiar e acessar esses dados, além de ser possível incorporá-los a outros sites ou programas.

2.4 Estudos de caso

Com o objetivo de demostrar as potencialidades do arcabouço computacional desenvolvido, foram realizados dois estudos de caso. O primeiro estudo de caso contém genomas da família das *Xanthomonadaceae*, sendo que inicialmente o conjunto selecionado

continha apenas 69 genomas que foram utilizados para validar o algoritmo de agrupamento de sequências. Uma vez validado o algoritmo, o restante do estudo de caso foi conduzido com um conjunto de 161 genomas (detalhados no Anexo A.3), pertencentes aos gêneros *Pseudoxanthomonas* (3), *Stenotrophomonas* (19), *Xanthomonas* (125) e *Xylella* (14). A escolha destes genomas se deve ao fato dos dois primeiros gêneros não se associarem a plantas, por outro lado o restante é estritamente fito-patogênico (com exceção de uma única espécie). O segundo estudo de caso foi realizado com 55 genomas da espécie *Streptococcus pyogenes*, um patógeno humano que causa uma ampla gama de doenças invasivas e não invasivas (detalhados nos Anexos A e B).

Todos os genomas de ambos os estudos de caso foram reanotados pelo serviço disponível no *Patric web service* (WATTAM et al., 2017) que é baseado no método RASTtk (BRETTIN et al., 2015).

3 Resultados

O arcabouço computacional desenvolvido no decorrer deste projeto possui as ferramentas para abarcar todo o processo de uma pesquisa sobre pan-genomas, desde códigos necessários para baixar os genomas, reanotá-los e, por fim, analisá-los. As ferramentas presentes estão disponíveis para acesso no *github* (em ⟨https://github.com/caiorns/GTACG-backend⟩) e podem ser executadas em computadores com sistema operacional Linux. Nas seções seguintes são descritos os resultados obtidos pelos experimentos realizados pelo arcabouço computacional, bem como a análise de desempenho de algumas das funcionalidades desenvolvidas.

3.1 Avaliação da identificação de famílias homólogas

Dois conjuntos de genomas foram utilizados com o objetivo de avaliar o agrupamento (ou clusterização) de sequências, o primeiro é composto por 69 genomas da família *Xanthomonadaceae* e o segundo possui 55 genomas da espécie *Streptococcus pyogenes*, ambos estão descritos no Anexo A. Todos os genomas passaram pela reanotação de suas CDS utilizando *Patric web service*, resultando em um total de, respectivamente, 309.147 e 101.220 CDS. Essas sequências foram agrupadas a fim de se encontrar famílias de genes, gerando um total de 48.477 e 4.466 famílias, respectivamente. Com o objetivo de analisar a qualidade do agrupamento de sequências foi realizado um experimento para comparar os resultados do algoritmo proposto com o TribeMCL. Este algoritmo foi escolhido por ser base para outros serviços (LI, 2003; PAGE et al., 2015; CONTRERAS-MOREIRA; VINUESA, 2013) e por ter sua qualidade reconhecida (BROHÉE; HELDEN, 2006). Os conjuntos tiveram suas sequências agrupadas pelos dois algoritmos. Os resultados das execuções foram avaliados seguindo um princípio de classificação de dados, da seguinte forma: as famílias encontradas pelos dois algoritmos foram utilizadas como o resultado da classificação e a anotação das sequências foi entendida como a resposta de fato esperada. Duas a duas, as sequências foram comparadas, caso estas sequências estejam na mesma componente conexa (mesma família) o resultado da classificação foi considerado Positivo (P), caso contrário foi considerado Negativo (N). Se as sequências possuem as mesmas funções anotadas, o resultado da classificação foi considerado Verdadeiro (V) e, caso contrário, Falso (F).

Essa abordagem permite analisar o problema por meio de métricas comuns a problemas de classificação, como:

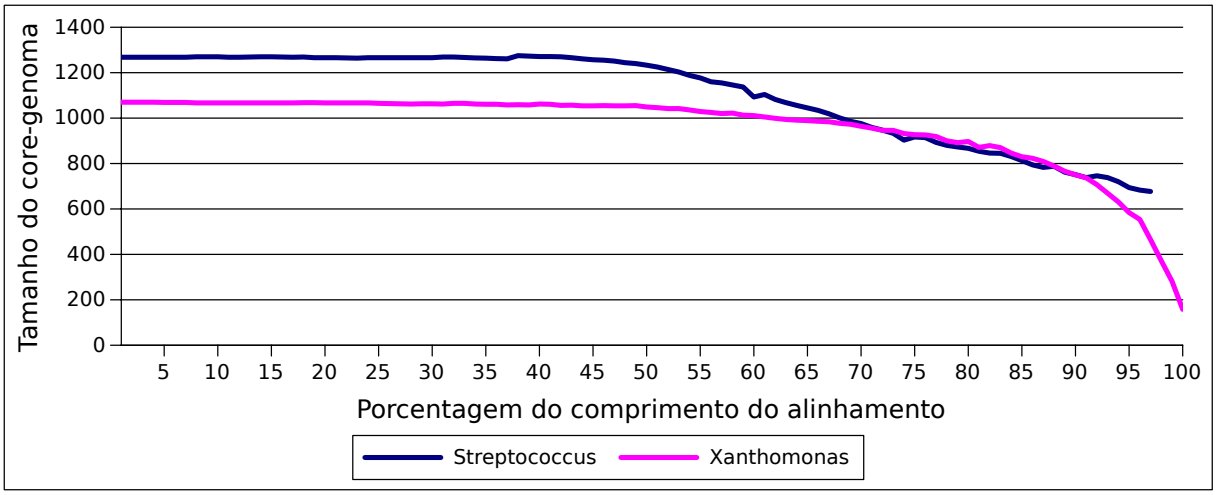
- Acurácia: $\frac{VP + VN}{VP + VN + FP + FN}$
- Sensibilidade: $\frac{VP}{VP + FP}$
- Especificidade: $\frac{VN}{VN + FN}$
- Eficiência: $\frac{Sensibilidade}{Especificidade}$

Diferente do que é comum a outros problemas de classificação, uma parte considerável das classes (neste caso as funções anotadas) são desconhecidas (entre 13% e 27%), e são marcadas como proteínas hipotéticas. Além disso, não há total confiança sobre a anotação, já que por ser um método automático ele está mais suscetível a falhas. Um especialista poderia fazer uma anotação mais confiável, mas com o entrave de ter um alto custo de tempo e esforço. Assim, destaca-se que essas limitações resultam em uma incerteza na avaliação da qualidade do agrupamento. Para permitir uma análise mais detalhada da avaliação, os resultados foram calculados e são exibidos em dois grupos, o primeiro considerando apenas os resultados utilizando as sequências com funções conhecidas (não hipotéticas) e o segundo considerando o conjunto como um todo.

Como os dois estudos de caso utilizam genomas com alta similaridade entre si (todos de mesma família ou de mesma espécie), é de se esperar que tenham um core-genoma composto por um número grande de famílias. Mantendo-se o e-value máximo fixo em 10^{-10}, o comprimento percentual do alinhamento foi variado de forma a se identificar qual valor maximizaria o tamanho do core-genoma. A figura 6 apresenta a distribuição do número de famílias no core-genoma em função da porcentagem do comprimento do alinhamento, tendo seu máximo em 38% para o conjunto *S. pyogenes*. Já para o conjunto *Xanthomonadaceae*, a função se mostrou decrescente, assim a porcentagem sobre o comprimento de alinhamento foi definida, empiricamente, como 30%. O agrupamento produziu, além dos resultados do alinhamento inicial, seis camadas de corte sobre o e-value para o conjunto *S. pyogenes* (10^{-14}, 10^{-27}, 10^{-43}, 10^{-46}, 10^{-47}, 10^{-51} e 10^{-59}), já para o conjunto *Xanthomonadaceaes* foram produzidas quatro camadas (10^{-15}, 10^{-23}, 10^{-31}, 10^{-35} e 10^{-46}). Após aplicadas as múltiplas camadas de corte, as famílias obtidas resultaram em core-genoma de 1.275

famílias para o conjunto de *S. pyogenes*, enquanto que o conjunto de *Xanthomonadaceaes* obteve um core-genoma de 1.063 famílias.

Figura 6 – Número de famílias no core-genoma utilizando o algoritmo de agrupamento Multilayer Clustering, nos estudos de caso com o conjunto de *Xanthomonadaceae* e *S. pyogenes*. O e-value foi mantido fixo em 10^{-10} e a porcentagem do comprimento do alinhamento variou no intervalo de 1 a 100.



Fonte: Santiago, Pereira e Digiampietri (2018)

A mesma estratégia foi empregada sobre o parâmetro *inflation* do TribeMCL, no entanto de forma exploratória. Assim, os resultados discutidos a seguir são baseados no uso do valor 15,0 para o parâmetro *inflation* para o grupo *S. pyogenes* e 10,0 para o grupo *Xanthomonadaceae*, totalizando 1.237 e 988 famílias no core-genoma, respectivamente.

Os resultados obtidos por ambos algoritmos analisados se mostraram bastante positivos, dada a natureza complexa do problema (Tabelas 1 e 2). Destaca-se que, para o algoritmo proposto, são apresentados os resultados para as diferentes camadas (ou níveis) de corte de valores de e-value. Porém, a discussão dos resultados considerará apenas o agrupamento final (isto é, os resultados produzidos na última camada).

O algoritmo proposto obteve resultados equivalentes ou melhores considerando a métrica de acurácia (Figuras 7 e 8). Por outro lado, considerando a medida Verdadeiros-Positivos (VP), o TribeMCL obteve valores mais altos para o grupo *S. pyogenes*.

As figuras 9 e 10 apresentam os resultados para a métrica de sensibilidade. Embora os resultados obtidos para a solução proposta sejam melhores para o conjunto de genomas filogeneticamente mais próximos (*S. pyogenes*), o mesmo não aconteceu com o conjunto mais distante (*Xanthomonadaceae*).

Tabela 1 – Resultados da classificação utilizando o algoritmo proposto

Camada	Classificação	Sem hipotéticos		Com hipotéticos	
		Streptococcus	*Xanthomonadaceae*	*Streptococcus*	*Xanthomonadaceae*
1ª	VP	2.610.724	17.215.957	2874088	18.163.658
	FP	3.281.104	31.994.843	3307026	34.657.608
	VN	3.874.748.818	25.176.769.293	5.105.439.320	46.961.418.375
	FN	7.083.285	63.462.660	11.073.156	768.757.312
2ª	VP	2.472.000	13.356.460	2.725.443	14.281.810
	FP	840.629	18.884.252	864.292	20.530.341
	VN	3.877.189.293	25.189.879.884	5.107.882.054	46.975.545.642
	FN	7.222.009	67.322.157	11.221.801	772.639.160
3ª	VP	2.462.605	12.340.091	2.715.989	13.258.950
	FP	793.497	9.861.963	817.075	10.606.057
	VN	3.877.236.425	25.198.902.173	5.107.929.271	46.985.469.926
	FN	7.231.404	68.338.526	11.231.255	773.662.020
4ª	VP	2.447.485	10.747.632	2.700.869	11.650.042
	FP	788.443	6.101.948	812.021	6.636.351
	VN	3.877.241.479	25.202.662.188	5.107.934.325	46.989.439.632
	FN	7.246.524	69.930.985	11.246.375	775.270.928
5ª	VP	2.439.807	–	2.693.069	–
	FP	410.250	–	433.731	–
	VN	3.877.619.672	–	5.108.312.615	–
	FN	7.254.202	–	5.108.312.615	–
6ª	VP	2.403.435	–	2.655.104	–
	FP	329.082	–	351.925	–
	VN	3.877.700.840	–	5.108.394.421	–
	FN	7.290.574	–	11.292.140	–

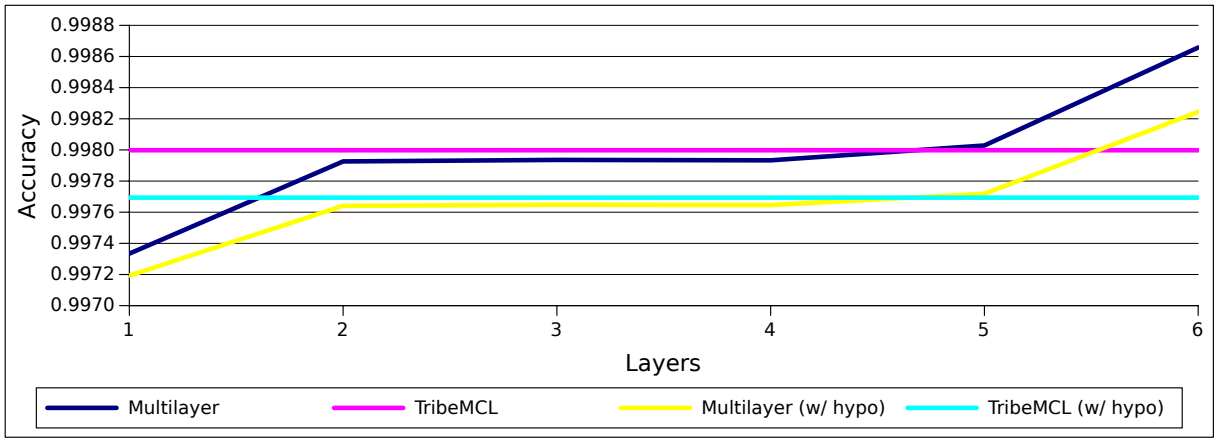
Fonte: Santiago, Pereira e Digiampietri (2018)

Tabela 2 – Resultado da classificação utilizando o TribeMCL

Classificação	Sem hipotéticos		Com hipotéticos	
	Streptococcus	*Xanthomonadaceae*	*Streptococcus*	*Xanthomonadaceae*
VP	2.510.553	8.804.005	2.787.488	9.773.633
FP	599.795	2.458.627	655.159	2.948.122
VN	3.877.430.127	25.206.305.509	5.108.091.187	46.993.127.861
FN	7.183.456	71.874.612	11.159.756	777.147.337

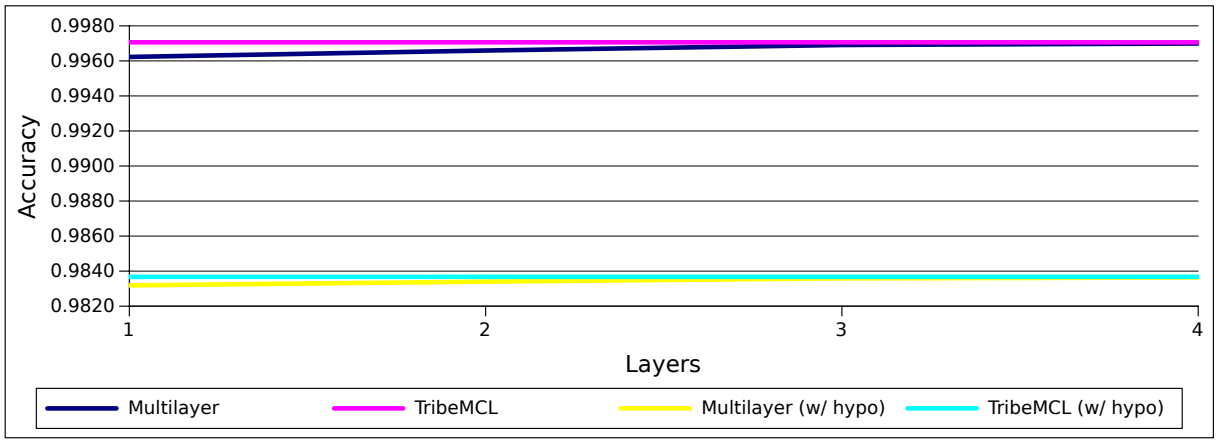
Fonte: Santiago, Pereira e Digiampietri (2018)

Figura 7 – Acurácia para o conjunto de genomas de *S. pyogenes*



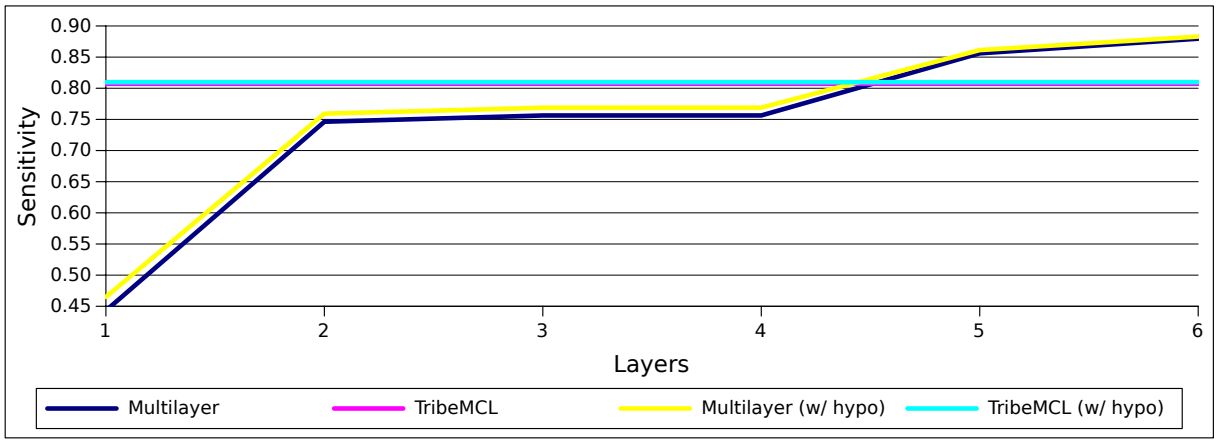
Fonte: Santiago, Pereira e Digiampietri (2018)

Figura 8 – Acurácia para o conjunto de genomas de the Xanthomonadaceae



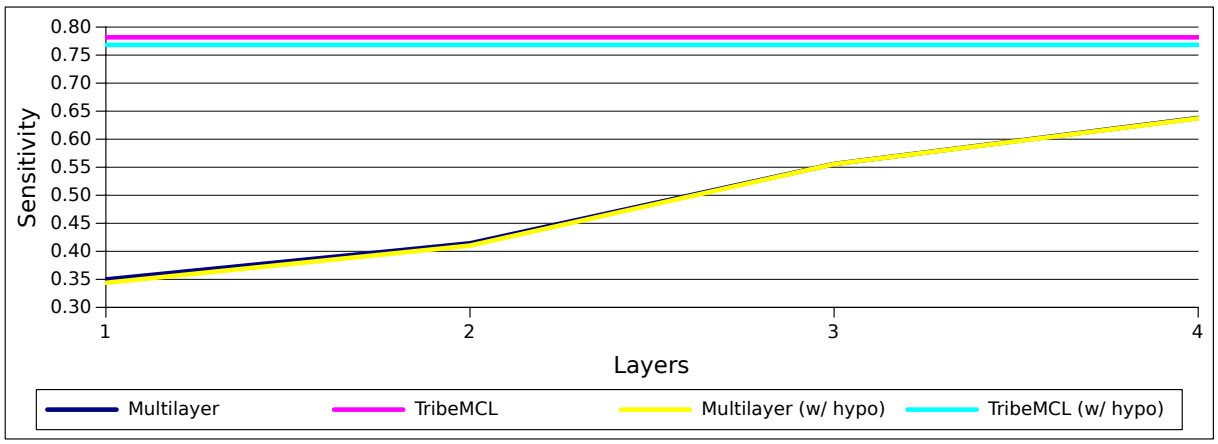
Fonte: Santiago, Pereira e Digiampietri (2018)

Figura 9 – Sensibilidade para o conjunto de genomas de Streptococcus pyogenes



Fonte: Santiago, Pereira e Digiampietri (2018)

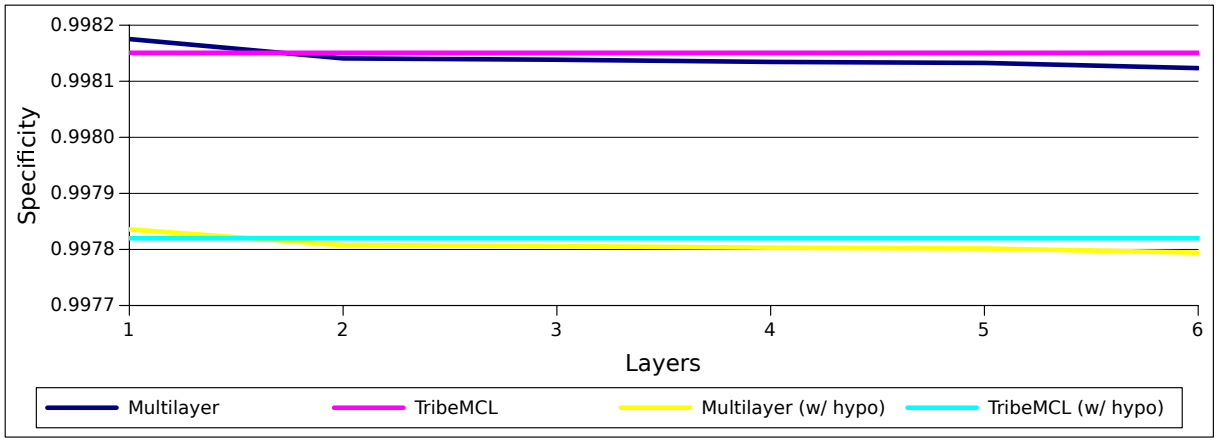
Figura 10 – Sensibilidade para o conjunto de genomas de Xanthomonadaceae



Fonte: Santiago, Pereira e Digiampietri (2018)

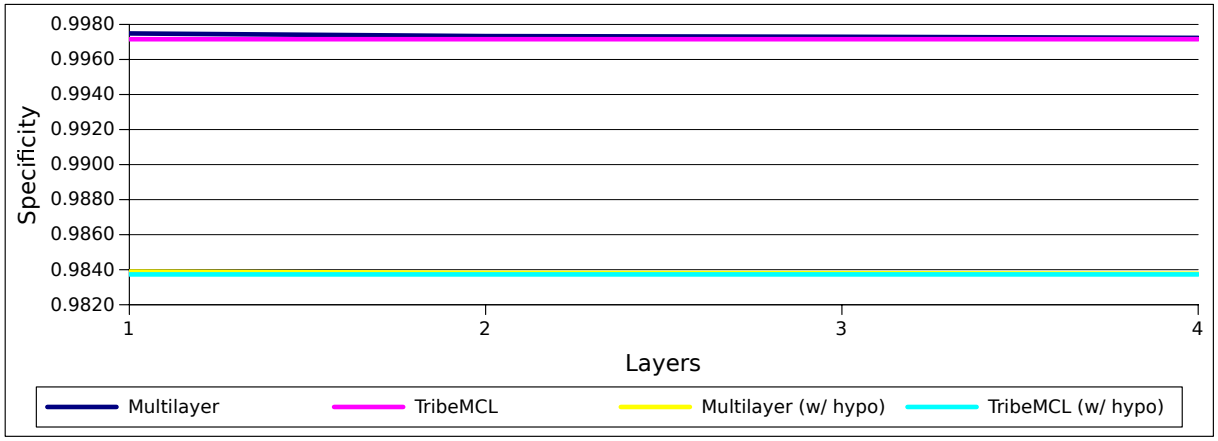
As diferenças encontradas para a métrica de especificidade (Figuras 11 e 12) entre os dois algoritmos foram pequenas (menos de 0,01%), por causa da grande quantidade de Verdadeiros-Negativos (VN).

Figura 11 – Especificidade para o conjunto de genomas de Streptococcus pyogenes



Fonte: Santiago, Pereira e Digiampietri (2018)

Figura 12 – Especificidade para o conjunto de genomas de Xanthomonadaceae

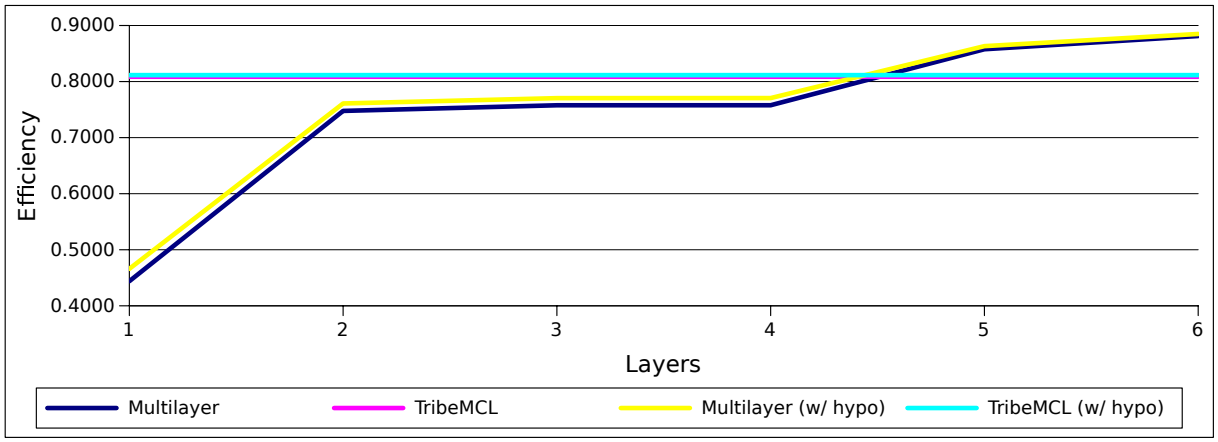


Fonte: Santiago, Pereira e Digiampietri (2018)

As questões levantadas pela métrica de sensibilidade são refletidas na eficiência de forma equivalente (Figuras 13 e 14), isto é, a solução proposta foi mais eficiente para o conjunto *S. pyogenes* do que o TribeMCL, o que não aconteceu com o conjunto *Xanthomonadaceae*.

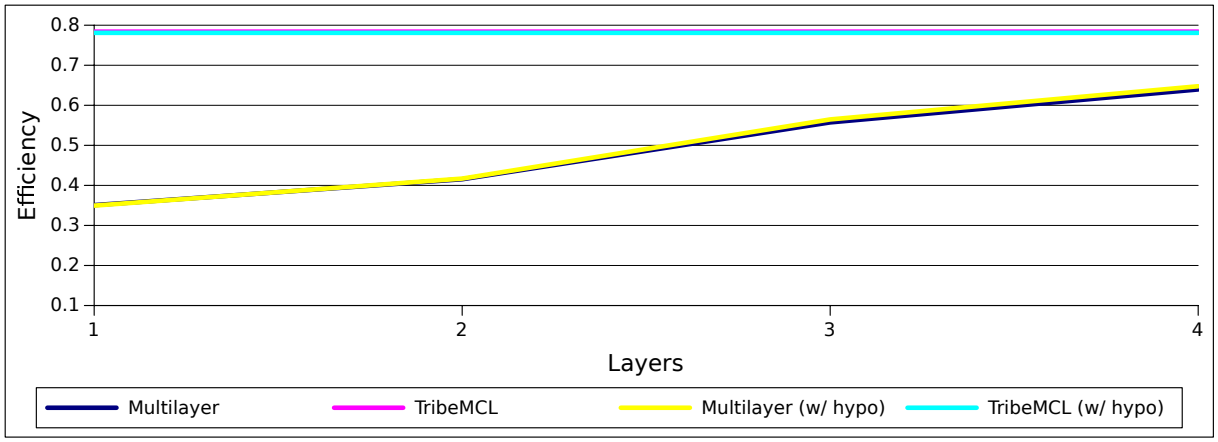
Como o algoritmo de agrupamento proposto preserva as relações entre os vértices do grafo, isso é, os alinhamentos podem ser descartados, mas nunca criados ou modificados, isso abre diversas possibilidades de análises topológicas. Uma delas é a identificação de

Figura 13 – Eficiência para o conjunto de genomas de Streptococcus pyogenes



Fonte: Santiago, Pereira e Digiampietri (2018)

Figura 14 – Eficiência para o conjunto de genomas de Xanthomonadaceae



Fonte: Santiago, Pereira e Digiampietri (2018)

possíveis domínios e *motifis*, que são muito relevantes para estudos genéticos (VOGEL et al., 2004).

A classificação dos domínios é diferente da realizada anteriormente para os grupos homólogos, pois um mesmo grupo pode ter sequências com diferentes domínios. Assim, os grupos não podem ser vistos como conjuntos disjuntos, apesar das funções anotadas continuarem tendo apenas um único valor. Para cada grupo foram calculados os valores de verdadeiro, falso, positivo e negativo, em que dado um domínio, os vértices que pertencem ao domínio são considerados positivos e os que não pertencem são considerados negativos. As tabelas 3 e 4 apresentam os resultados desta análise.

A acurácia e a especificidade variaram menos de 0,5% em relação aos resultados anteriores. Porém, o principal avanço foi na métrica de sensibilidade e isso se reflete direta-

Tabela 3 – Resultados da classificação utilizando o algoritmo de identificação multidomínios

Classificação	Sem hipotéticos		Com hipotéticos	
	Streptococcus	Xanthomonadaceae	Streptococcus	Xanthomonadaceae
VP	1.898.096	4.313.358	2.655.104	5.058.857
FP	200.631	463.626	351.925	643.403
VN	4.890.956.187	4.374.1697.616	5.108.394.421	75.931.484.135
FN	8.841.324	293.378.177	11.292.140	1.041.731.867
Acurácia	0,9982	0,9933	0,9977	0,9865
Sensibilidade	0,9044	0,9029	0,8830	0,8872
Especificidade	0,9982	0,9933	0,9978	0,9865
Eficiência	0,9060	0,9090	0,8849	0,8993

Fonte: Santiago, Pereira e Digiampietri (2018)

Tabela 4 – Acurácia da classificação considerando a identificação multidomínios

Conjunto	Algoritmo	Sem hipotéticos		Com hipotéticos	
		Sem domínios	Com domínios	Sem domínios	Com domínios
Streptococcus	Multilayer	0,8795681783	0,9044034789	0,8829658776	0,8829658776
	TribeMCL	0,8071614495	0,8071614495	0,9978200412	0,9978200412
Xanthomonadaceae	Multilayer	0,6378575608	0,9029458755	0,9837689600	0,8871670180
	TribeMCL	0,7817004942	0,7817004942	0,9837315709	0,9837315709

Fonte: Santiago, Pereira e Digiampietri (2018)

mente na eficiência do algoritmo. O grupo formado por genomas da espécie *Streptococcus pyogenes* teve sua sensibilidade aumentada de 87,9% para 90,4%. Já o aumento de sensibilidade do grupo formado por genomas da família *Xanthomonadaceae* foi consideravelmente maior, de 63,7% para 90,9%. Essas mudanças causaram um impacto direto na eficiência da classificação que melhorou de 88,1% para 90,6% e 63,9% para 90%, respectivamente.

3.2 Visualização de resultados

Os resultados visuais foram estruturados no formato de um *website* estático com diferentes níveis de detalhamento. Na tela inicial do *website* estão contidas as informações mais macroscópicas. Este primeiro nível está relacionado a visualização e interação com dados genômicos (Figura 15). Esta tela está dividida em cinco seções, nas duas primeiras (*Settings* e *Filters*, Figuras 15A e 15B) estão os meios pelos quais o usuário tem acesso aos níveis seguintes de detalhamento dos dados, em que são listadas as famílias de acordo com métricas estatísticas. Na seção Filters é possível definir critérios para filtrar as famílias que são listadas no nível seguinte, exigindo a ausência ou a presença dos genomas de forma individual ou dos grupos de genomas (além dos grupos de genomas definidos na fase de pré-processamento, também é possível criar grupos em tempo real com o inconveniente de nem todas as métricas estarem disponíveis). Por meio de filtros encontram-se, por exemplo, todas as sequências que são compartilhadas apenas por determinado grupo de

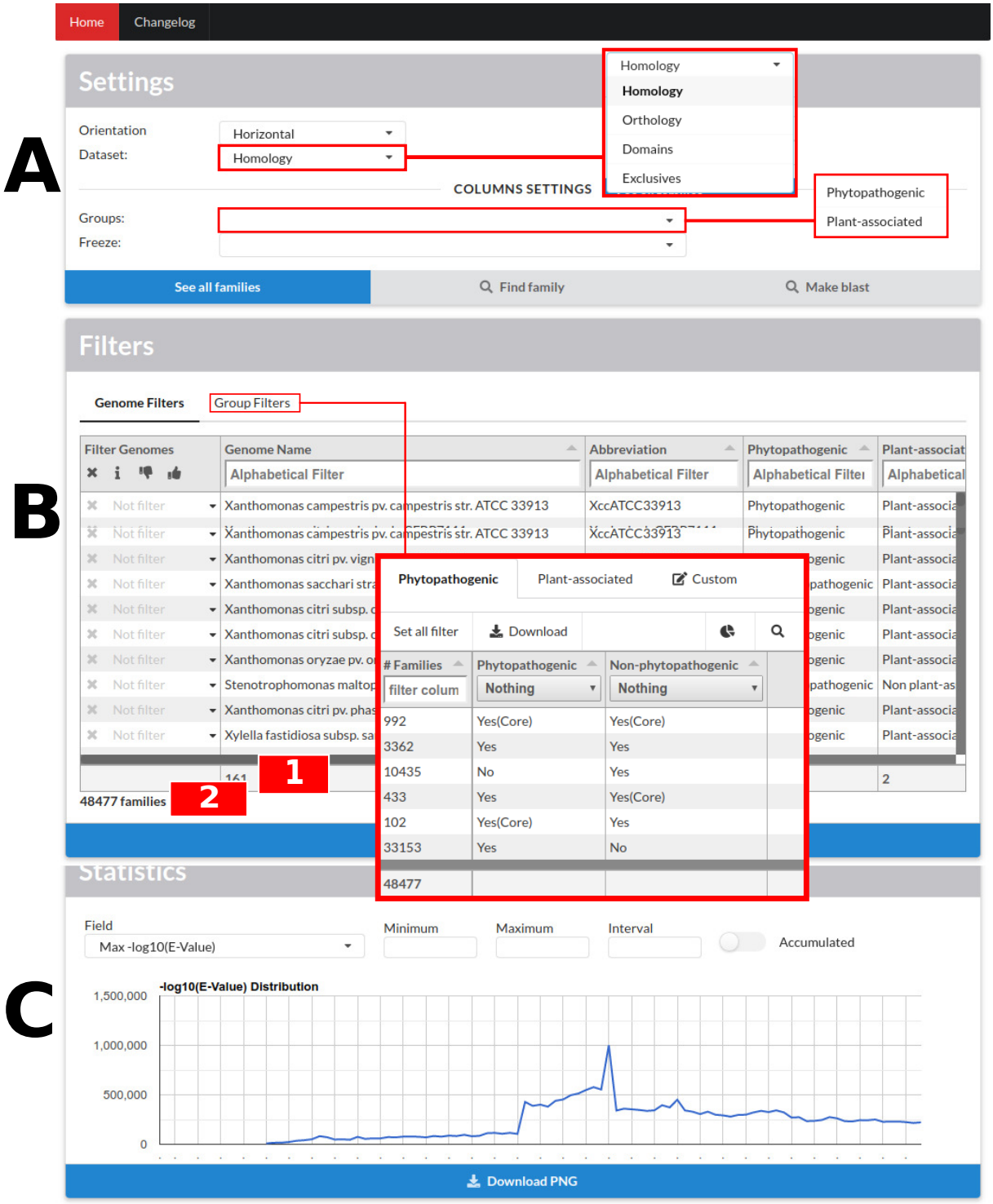
genomas. Na seção seguinte (*Statistics*, Figura 15C) são apresentados gráficos, produzidos utilizando a biblioteca Google Charts, baseados em métricas sobre famílias, sequências e alinhamentos locais. Por fim, as seções seguintes (*2D Plot* e *Phylogeny*, Figuras 15D e 15E) apresentam os métodos escolhidos para visualização de genomas. Essas duas seções podem ser customizadas com base nos grupos de genomas e outras configurações adicionais. A filogenia é exibida utilizando a biblioteca Phylocanvas.

O nível seguinte é destinado ao estudo estatístico das famílias. Neste nível, famílias podem ser pesquisadas por meio de métricas como: o número de genomas que as compartilham, número de sequências, distribuição do comprimento das sequências, função anotada, métricas baseadas nos grafos, métricas baseadas nos alinhamentos, métricas baseadas nas filogenias e métricas baseadas nas anotações manuais dos grupos de genomas.

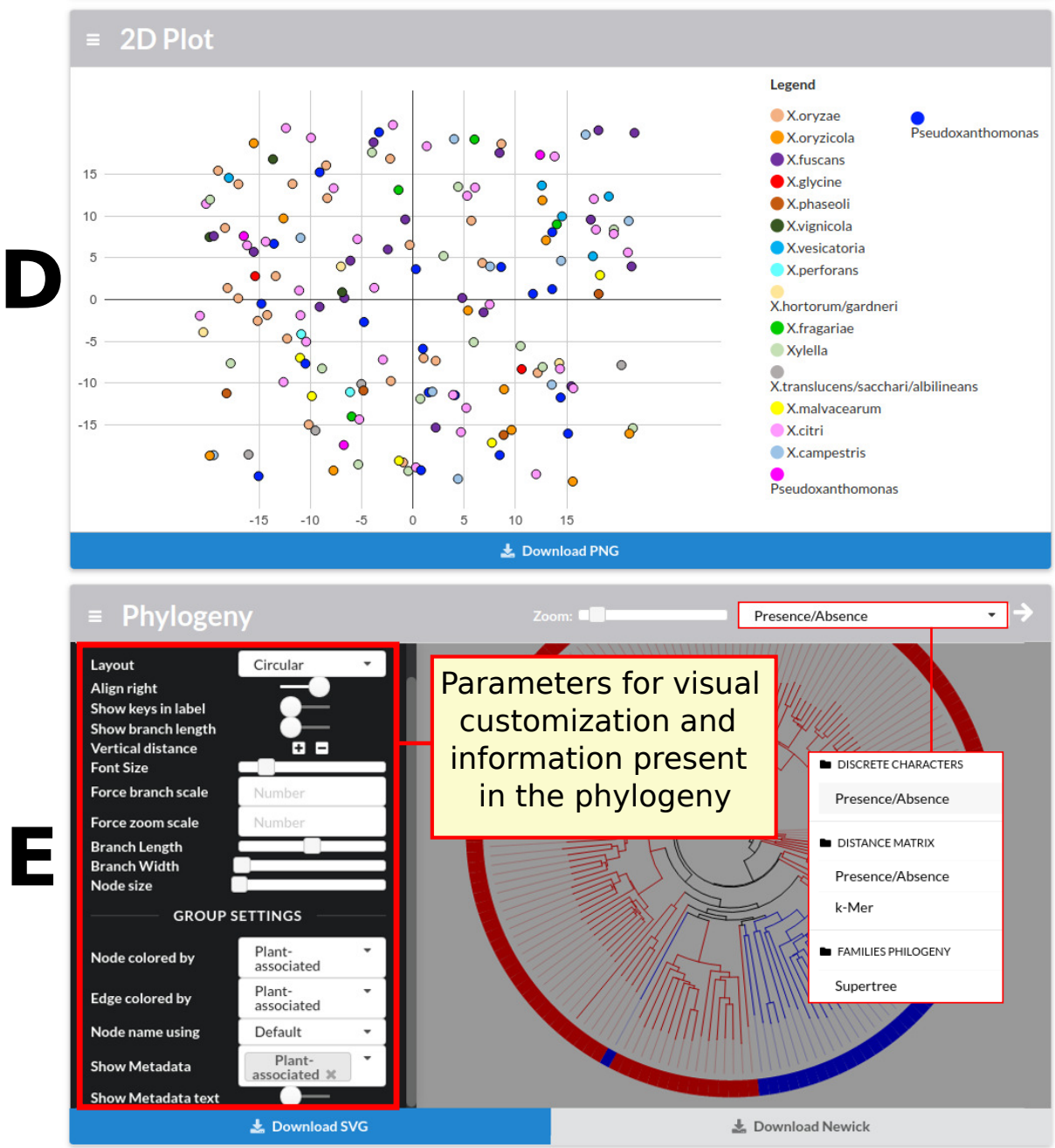
Os dados estatísticos visualizados são específicos para a granularidade das subdivisões das famílias (homologia, ortologia e domínios já discutidos anteriormente), escolhida pelo usuário na tela inicial. Esses dados estão disponíveis para serem baixados em formatos que podem ser usados para a construção de uma filogenia (uma matriz de distância, por exemplo) ou no formato utilizado pelo Roary (PAGE et al., 2015), abrindo uma ampla gama de funções para análise e visualização de dados. Em relação às sequências, famílias podem ser encontradas de acordo com as métricas associadas às sequências que compõem cada família, como a função anotada, comprimento, ou posição no genoma. Por meio de uma simples configuração no servidor (isto é, a execução de um código escrito em *Node.js*) é possível habilitar no arcabouço a busca por famílias por meio da ferramenta BLAST contra as sequências do pan-genoma, utilizando os filtros e resultados já discutidos anteriormente. Essa abordagem pode ser feita com as tabelas dinâmicas fornecidas pela biblioteca Tabulator, deste modo o usuário tem a sua disposição filtros dinâmicos e complexos que podem utilizar expressões lógicas e matemáticas, além de permitir o agrupamento de dados.

O último e mais baixo nível de detalhamento do sistema é relativo às famílias. Neste nível, cada família tem uma página com suas respectivas informações (Figure 16). Assim como a página inicial, este nível é dividido em cinco seções. A primeira seção contém informações sobre as sequências (anotação, comprimento, entre outras), combinada com as informações de seus respectivos genomas (identificação do genoma e a anotação dos grupos). Para cada sequência também está presente um *link* de acesso a uma busca do BLAST contra a base de aminoácidos do NCBI. Caso haja uma configuração básica

Figura 15 – Tela inicial do GTACG. Estes resultados estão divididos em cinco seções: Settings, Filters, Statistics, 2D Plot, and Phylogeny. As duas primeiras são referentes a buscas subsequentes sobre as famílias. (C) Na terceira são apresentados gráficos sobre métricas referentes a famílias, sequências e alinhamentos locais. (D) A quarta apresenta a projeção bidimensional dos genomas. (E) Por fim, a última apresenta as filogenias construídas e opções de customização.



(Continuação)



Fonte: Santiago et al. (2019)

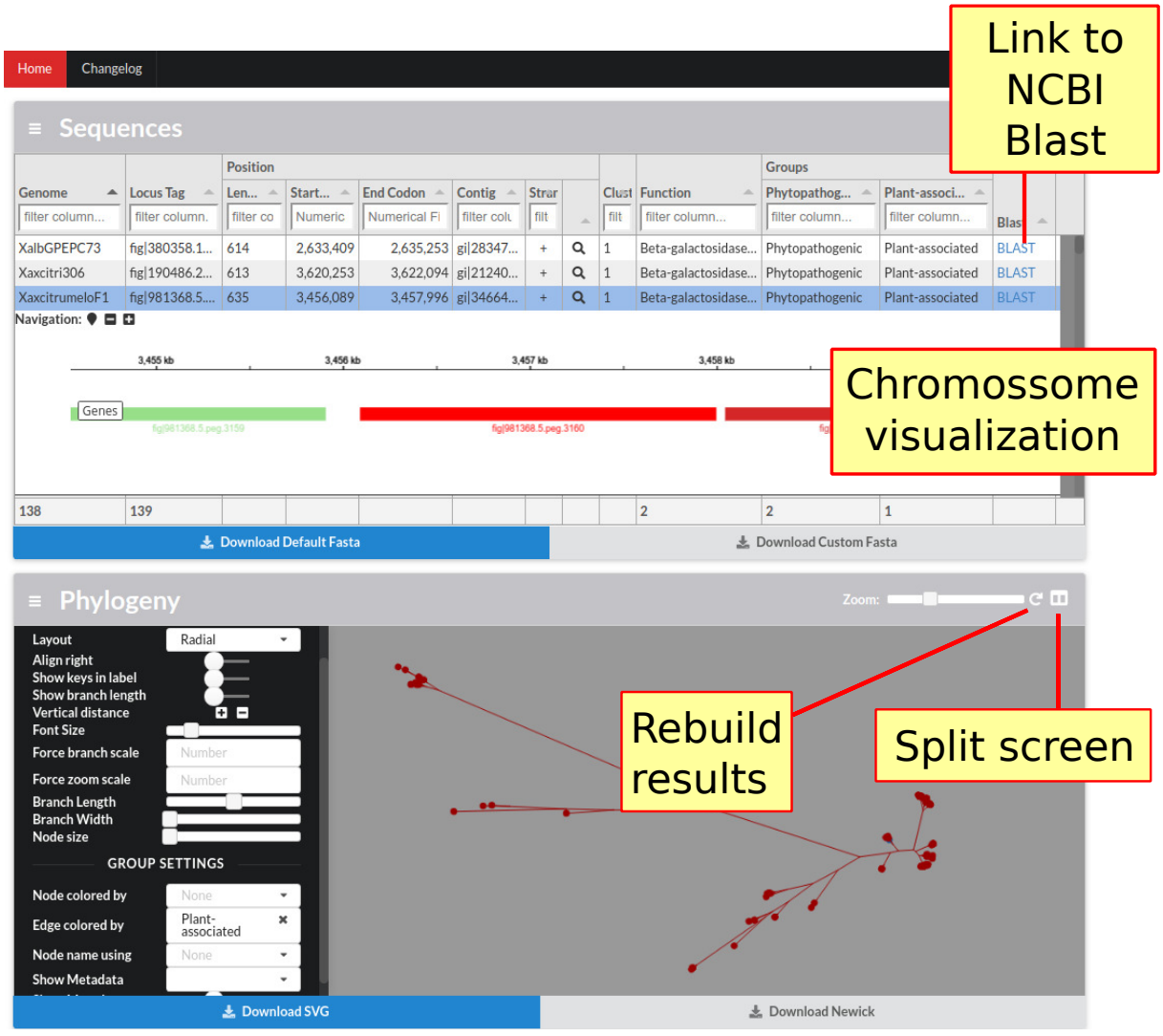
de servidor (um código escrito em Node.js), também é possível visualizar a sequência escolhida, assim como a sua sintenia, graça a biblioteca igv.js. Nas duas seções seguintes estão a visualização da filogenia e do alinhamento das sequências, construídas a partir das bibliotecas Phylocanvas e MSAViewer (YACHDAV et al., 2016). Estes resultados foram calculados na fase de pré-processamento, porém com o servidor configurado (um código escrito em Node.js) esses resultados podem ser recalculados usando outros programas e parâmetros, como o FastTree (PRICE; DEHAL; ARKIN, 2010), o PhyML (GUINDON et al., 2010), o RaxML (STAMATAKIS, 2014), o Clustal Omega (SIEVERS et al., 2011) e o MUSCLE (EDGAR, 2004).

A quarta seção é destinada a apresentação do grafo gerado para a família, durante o processo de identificação de famílias. As sequências são representadas como vértices e os alinhamentos locais são representados como arestas. O grafo é exibido utilizando a biblioteca Sigma.js. As funcionalidades desenvolvidas permitem que o usuário investigue a situação dos alinhamentos das famílias, destacando alinhamentos de acordo com condições definidas, por exemplo, destacando todos os alinhamentos que possuem identidade menor que 80%. Por fim, a última seção contém uma sumarização estatística das métricas relativas aos grupos de genomas da família em questão (que foram definidos inicialmente na fase de pré-processamento).

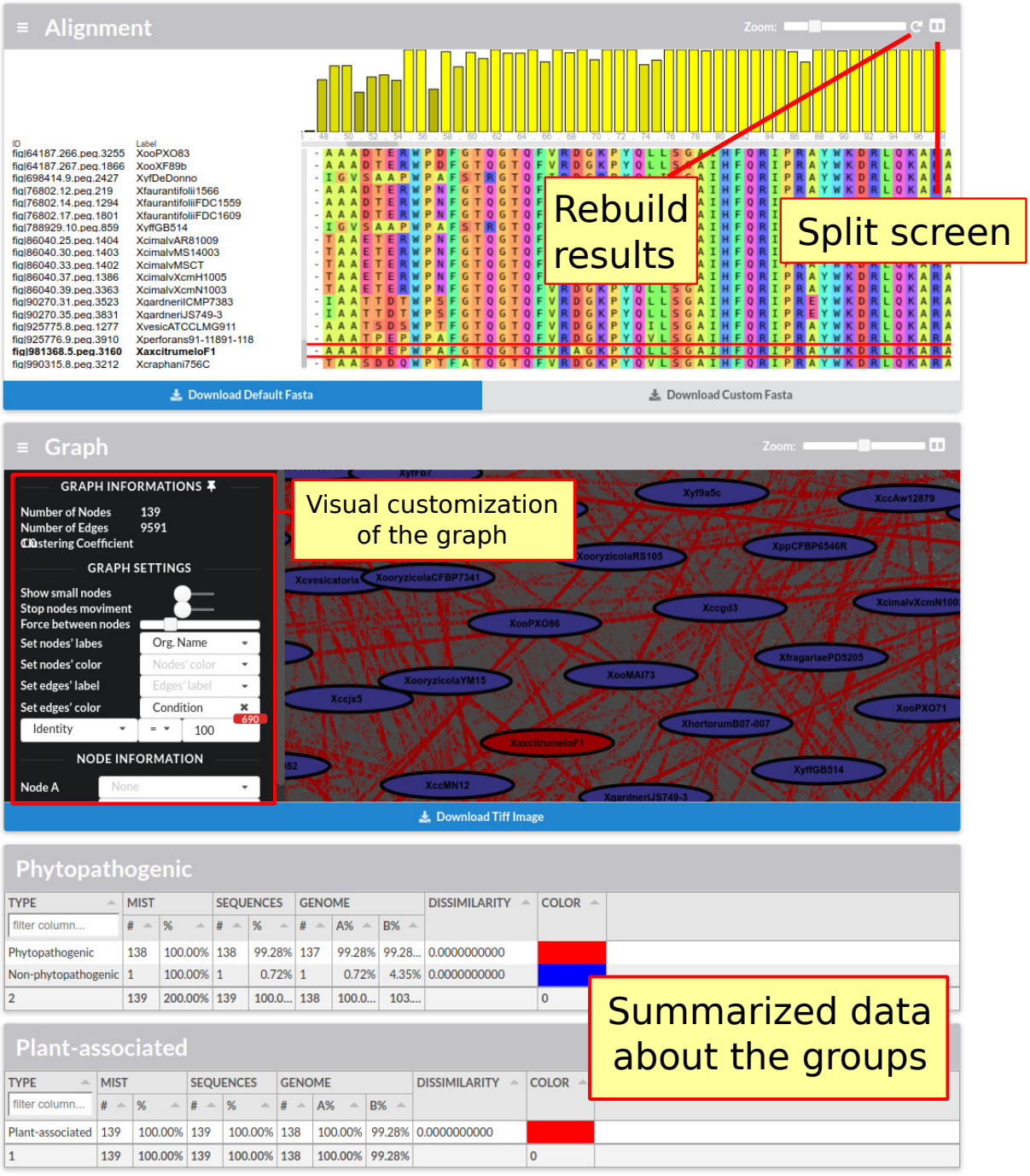
As funcionalidades disponibilizadas por este arcabouço computacional permitem ao usuário estruturar uma pesquisa utilizando uma abordagem *top-down*, começando com dados genômicos (como anotações fenotípicas, filogenias ou um levantamento de genes exclusivos, por exemplo) para então fazer uma investigação minuciosa mais profunda para entender os mecanismos genéticos que podem justificar os dados iniciais. O processo também pode ser invertido, os usuários podem partir de sequências de aminoácidos para encontrar a respectiva família e verificar diferentes informações dessa família no contexto dos grupos anotados de genomas. Para auxiliar os usuários finais, os resultados gráficos à disposição do usuário podem ser exportados em formatos com qualidade adequada à publicação, como SVG, TIFF e PNG.

Pelo fato de se tratar de um *website*, o compartilhamento de resultados e buscas é simplificado, pois basta que se copie a URL de determinada pesquisa para que os estados estabelecidos durante a navegação sejam compartilhados para um trabalho em equipe. E, uma vez que os dados já tenham sido produzidos, não há necessidade de nenhum tipo de instalação para se usufruir dos benefícios trazidos por este arcabouço computacional.

Figura 16 – Tela referente a uma família. As informações contidas nesta tela estão organizadas em quatro principais seções, seguidas de uma sumarização das informações sobre os grupos de genomas relativos à família em questão. (A) A primeira seção contém dados sobre as sequências e seus respectivos genomas; e caso haja uma configuração de servidor, é possível visualizar as sequências de forma posicional em conjunto com sua vizinhança. (B) Na segunda seção, é possível visualizar, customizar e reconstruir (com diferentes parâmetros) a filogenia das sequências. (C) Na seção seguinte, é possível visualizar, customizar e reconstruir (com diferentes parâmetros) o alinhamento das sequências. (D) E finalmente, a última seção apresenta o grafo construído na etapa de identificação das famílias, em que as sequências são representadas como vértices e os alinhamentos são representados como arestas. O grafo pode ser personalizado para destacar alinhamentos de acordo com alguma métrica específica. Nesta figura, os alinhamentos locais com identidade menor que 98,5% estão destacados.



(Continuação)



Fonte: Santiago et al. (2019)

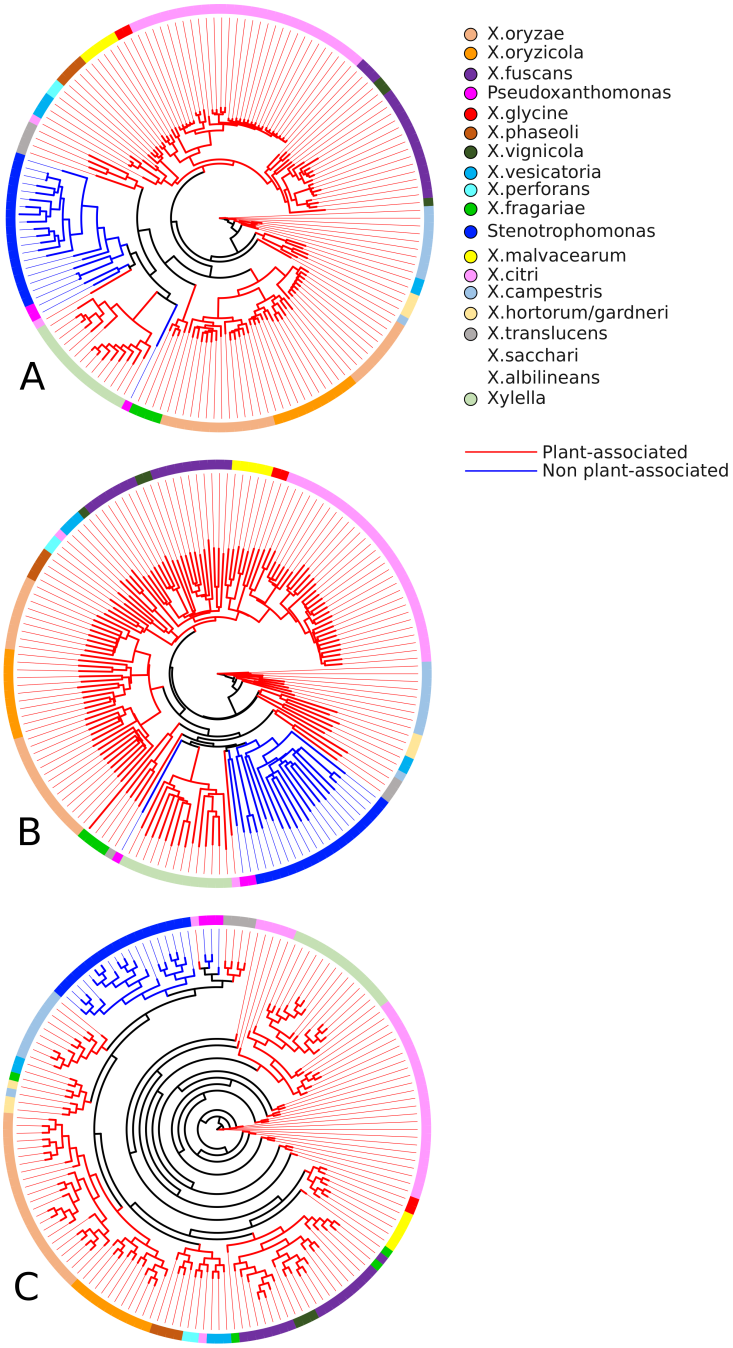
3.3 Estudos de casos

Os 161 genomas provenientes da família *Xanthomonadaceae* abordados neste estudo têm tamanho entre 2,5 e 5,5 milhões de pares de bases. A anotação automática desses genomas identificou uma média de 4.620 CDS por genoma, totalizando 743.920 CDS, as quais foram agrupadas em 48.477 famílias homólogas. Destas famílias, 4.287 foram subdivididas em 13.528 famílias ortólogas, resultando em um total de 57.718 famílias ortólogas. Essa quantidade de famílias ortólogas era esperada ao se considerar a complexidade e o tamanho desse conjunto de genomas. Para obter esses resultados foram definidos dois parâmetros: (1) um limiar máximo para o e-value em 10^{-10} e (2) um limiar mínimo de 45% para o comprimento do alinhamento.

Para este estudo de caso, o principal fenótipo de interesse para a avaliação do GTACG é associado ao fato de alguns microrganismos de gêneros específicos pertencentes à família *Xanthomonadaceae* possuírem uma associação adaptativa com plantas, quer como fitopatógenos ou não. É importante ressaltar que esta característica não é mandatória para todos os genomas deste conjunto. Ao todo são 139 genomas que apresentam essa característica, pertencentes aos gêneros *Xanthomonas* e *Xylella*, por outro lado, os 22 genomas dos gêneros *Pseudoxanthomonas* e *Stenotrophomonas* não apresentam essa característica.

Pelos métodos de inferência filogenética utilizados (Figura 17) fica clara a boa separação dos genomas dos organismos associados a plantas dos demais, tal como já foi relatado na literatura (SHARMA; PATIL, 2011). Nas filogenias construídas com base no vetor de características binárias (Figura 17A) e com base na matriz de distância (Figura 17B), é possível ver essa clara separação entre os grupos. Duas exceções estão presentes em ambas as árvores, o genoma *P. spadix* BD-a59 que foi agrupado junto com genomas associados a plantas, e o agrupamento de *X. mangiferaeindicae* junto com genomas que não são associados a plantas. Entretanto, a *supertree* (Figura 17C) apresentou um agrupamento com mais ancestrais hipotéticos do grupo de não associados a plantas, excluindo, portanto, a *Xylella* (em desacordo com as duas filogenias anteriores). Este resultado é corroborado com outros estudos que mostram que *Stenotrophomonas* é filogeneticamente mais próximo de *X. campestris* do que da *Xylella* (NAUSHAD; GUPTA, 2013; RAMOS et al., 2011).

Figura 17 – Filogenias estabelecidas pelo arcabouço para os conjuntos de genomas da família *Xanthomonadaceae*. A filogenia A foi inferida a partir dos vetores binários de características de cada genoma; as posições do vetor representam as famílias e são definidas como 0 ou 1, dependendo se o genoma possui ou não uma de suas sequências na família; para a inferência foi utilizado o programa de parcimônia (*pars*) para características binárias incluso no Phylip. A filogenia B foi construída utilizando a matriz de distância, calculada com base na distância euclidiana dos vetores de características binárias; o método escolhido foi o *neighbor-joining* presente no Phylip. A filogenia C foi construída pelo método da supertree, que sumariza todas as árvores filogenéticas construídas para as famílias; o método escolhido foi o *Quartet fit* com o *Nearest Neighbour Interchange* disponibilizada pelo Clann.



Fonte: Santiago et al. (2019)

Nenhuma família ortóloga possui o comportamento "ideal" (em termos de separação de grupos) de ser compartilhada por todos os genomas associados a plantas, e ao mesmo tempo não estar presente nos demais. Porém foram encontrados resultados interessantes e que são consistentes com a filogenia encontrada. Foram encontradas 19 famílias de genes compartilhadas por ao menos 90% dos genomas associados a plantas e ausentes a todos os outros. Destaca-se que esses genomas ausentes são os mesmos identificados como um grupo separado na filogenia. Em nenhuma dessas 19 famílias, a *X. mangiferaeindicae* está presente. Em três dessas famílias, a *X. albilineans* também não está presente, e em duas famílias, duas cepas de *X. translucens* e *X. sacchari* não estão presentes.

Também foram encontradas nove famílias que são compartilhadas por todos os genomas associados a plantas e por menos de 30% dos genomas não associados. De forma similar, os genomas não associados presentes nesse resultado se mostram integrados aos genomas associados de acordo com filogenia da *supertree*. A respeito destas nove famílias, o número de genomas não associados a plantas é relativamente pequeno (entre três e seis genomas). Este resultado era particularmente esperado, dados os resultados apresentados pela *supertree*, indicando que *P. spadix* BD-a59, *P. suwonensis* 11-1, e *P. suwonensis* J1 (que estão presentes nessas famílias) compartilham um mesmo ancestral hipotético recente com os genomas associados a plantas.

Duas famílias que compõem o core-genoma têm dissimilaridade maior que 1% em seus alinhamentos, o que indica a existência de bases específicas com mutações relacionadas ao grupo de genomas associados a plantas. Por fim, existem 13 famílias também do core-genoma que separam em um único ramo da filogenia todos os organismos associados a plantas.

Considerando o estudo de caso contendo os genomas de *Streptococcus pyogenes*, ao todo foram analisados 55 genomas com tamanho entre 1,7 e 2 milhões de pares de bases. A anotação automática identificou um total de 101.220 CDS (média de cerca de 2.250 CDS por genoma). Estas CDS foram agrupadas em 4.466 famílias homólogas, das quais 227 foram subdivididas em 514 famílias ortólogas, resultando em um total de 4.753 famílias ortólogas. Esse número é relativamente menor do que o obtido para o conjunto da família *Xanthomonadaceae*, não só pelo fato de ser um conjunto menor (com menos genomas e de tamanhos menores), mas por se tratar de um conjunto de menor diversidade genética, uma vez que engloba apenas genomas de uma mesma espécie. Para obter esses resultados

foram definidos dois parâmetros: (1) um limiar máximo do e-value em 10^{-10} e (2) um limiar mínimo de 41% do comprimento do alinhamento.

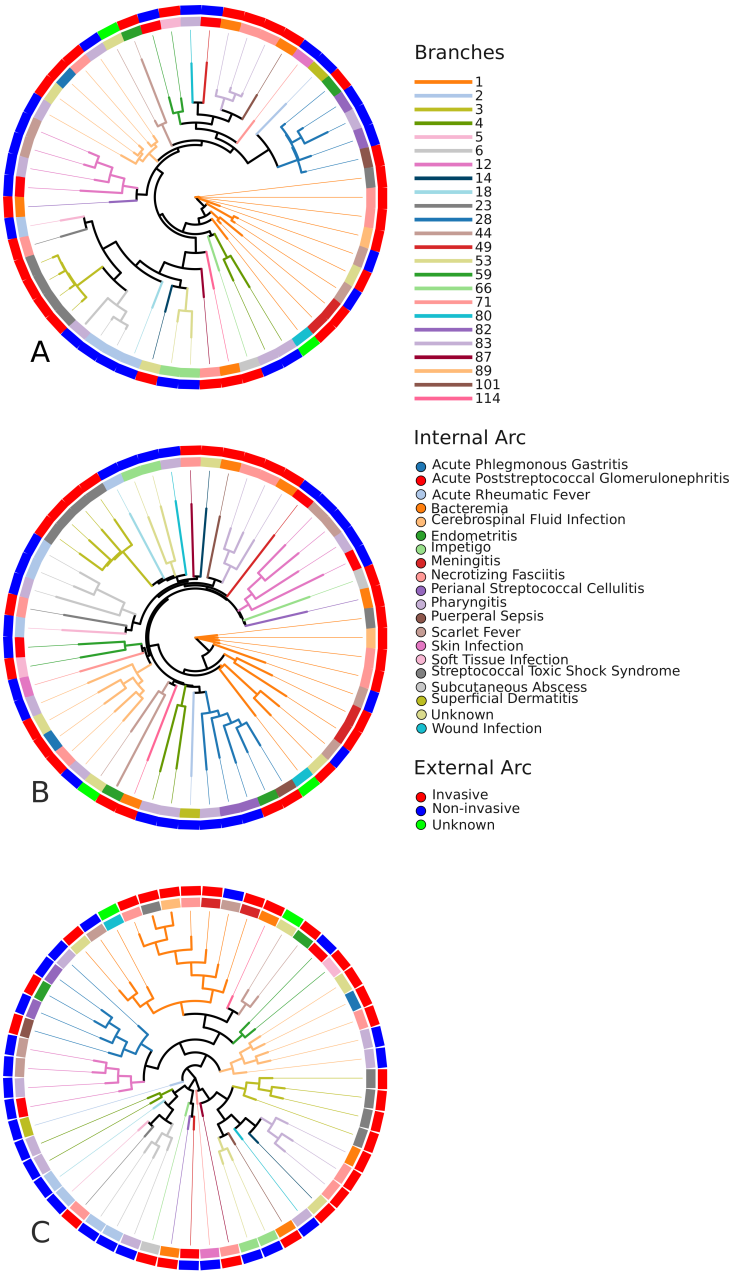
Neste conjunto existe mais de um fenótipo de interesse. O primeiro é referente a doenças causadas por esses microrganismos. O *S. pyogenes* é um patógeno humano capaz de causar uma ampla gama de doenças, desde simples faringites até infecções mais severas como fascite necrotizante ou bacteremia (BREIMAN et al., 1993; CUNNINGHAM, 2000; LAMAGNI et al., 2008). Dos 55 genomas estudados, foram anotados 20 tipos de doenças, coletadas de pacientes humanos que apresentaram essas infecções. Diretamente relacionado às doenças está a severidade dessas doenças, e para isso os genomas foram anotados com base na invasividade apresentada pela infecção, sendo anotados como invasivos, não invasivos ou de invasividade desconhecida. A descoberta de mecanismos genéticos envolvidos tanto na expressão das doenças quanto da invasividade delas pode ser de grande valia para novos estudos sobre vacinas ou antibióticos dedicados ao tratamento de infecções causadas pelo *S. pyogenes*. Porém, vale destacar que é possível que uma mesma bactéria seja capaz de causar diferentes doenças (de acordo com o local da infecção ou de características do paciente), mas, neste estudo de caso, cada genoma foi anotado de acordo com a doença que estava causando no paciente do qual o organismo foi isolado e posteriormente sequenciado.

Por fim, o último fenótipo de interesse é a expressão da proteína *emm*, uma região de alta variação genética e que é considerada um dos principais fatores de virulência para o *S. pyogenes* (LANCEFIELD; PERLMANN, 1952). Estudos epidemiológicos já mostraram a relação entre esta proteína com a patogenicidade do organismo (CUNNINGHAM, 2000; CARAPETIS et al., 2005; ENELI; DAVIES, 2007; SAKATA, 2013; TAMAYO et al., 2016).

Na figura 18 pode se notar que o genótipo *emm* é bastante correlato com a filogenia, as três árvores (baseadas em vetor de características binário, matriz de distância e *supertree*) conseguiram isolar todos os grupos de genomas de forma impecável. Isso corrobora o emprego dessa região do genoma para a classificação dos genomas dentro do *Group A Streptococcus* (LANCEFIELD; PERLMANN, 1952). Conforme esperado, existe uma grande correlação desses grupos com o ganho e perda de genes, assim como indicado pela quantidade de famílias de genes exclusivos para cada um desses grupos de genomas (Anexo C, Tabela C.1).

Nenhuma família conseguiu refletir exatamente a filogenia da região da proteína *emm*, por desafios ligados à anotação e ao agrupamento dessas sequências (muito em

Figura 18 – Filogenias estabelecidas pelo arcabouço para os conjuntos de genomas de *S. pyogenes*. A filogenia A foi inferida a partir dos vetores binários de características de cada genoma; as posições do vetor representam as famílias e são definidas como 0 ou 1, dependendo se o genoma possui ou não uma de suas sequências na família; para a inferência foi utilizado o programa de parcimônia (*pars*) para características binárias incluso no Phylip. A filogenia B foi construída utilizando a matriz de distância, calculada com base na distância euclidiana dos vetores de características binárias; o método escolhido foi o *neighbor-joining* presente no Phylip. A filogenia C foi construída pelo método da supertree, que sumariza todas as árvores filogenéticas construídas para as famílias; o método escolhido foi o *Quartet fit* com o *Nearest Neighbour Interchange* disponibilizada pelo Clann.



Fonte: Caio Santiago, 2019

parte por se tratar de região de alta diversidade). Porém, foram encontradas outras 15 famílias que desempenharam o mesmo papel de dividir esses grupos nos ramos filogenéticos esperados (de acordo com a literatura correlata). Este resultado permite a substituição deste marcador filogenético (*emm*), por ser mais fácil de ser obtido. Para conjuntos de genomas que possuem esta característica (apresentam famílias de genes capazes de separar filogeneticamente os genomas de acordo com alguma característica indicada pelo usuário, seja ela fenotípica ou genotípica) o arcabouço apresentado nesta tese permite a fácil identificação dessas famílias, sendo necessário apenas filtrar os grupos que tenham o valor 100% na métrica MIST.

Considerando os outros fenótipos estudados, não foi identificada a mesma relação filogenética, como pode ser visto pelo arco interno (sobre as doenças) e externo (sobre a invasividade) das árvores filogenéticas (Figura 18). Nenhuma família aparenta estar diretamente relacionada com a invasividade. O mesmo comportamento é visto no estudo sobre as doenças. A maior parte dos resultados que conseguem realizar bem parte da separação está diretamente relacionada aos grupos que são formados por genomas unitários (Anexo B.1, Tabela C.2). Tirando estes grupos, três famílias foram exclusivamente encontradas nos genomas anotados com impetigo (a toxina esfoliativa A e duas proteína hipotéticas), e outras quatro foram encontrados nas famílias exclusivas à meningite (um regulador transcricional, a proteína FtsQ de divisão celular e duas proteínas associadas a fagos). Entretanto, se tratam de grupos de baixa amostragem (apenas dois genomas cada), assim, os resultados precisam ser melhor investigados na literatura, além de ser repetidos com conjuntos mais representativos de genomas.

Quanto aos alinhamentos e às filogenias, os resultados são similares. As diferenças mais acentuadas (envolvendo mais de 20% das bases) se deram em famílias pequenas (com menos de 10 sequências). Nenhum resultado expressivo foi obtido quando a análise foi limitada ao core-genoma.

4 Discussão

Neste capítulo são discutidos os diferentes aspectos do arcabouço desenvolvido no presente projeto, organizados em seis seções.

4.1 Identificação de genes homólogos

O agrupamento de genes em famílias é uma tarefa complexa que visa a inferir relações de homologia com base em outras medidas, como as calculadas a partir de alinhamentos locais. Os resultados obtidos se mostraram positivos para ambos os algoritmos comparados: o desenvolvido no decorrer deste projeto, baseado no coeficiente de agrupamento e a ferramenta TribeMCL. Devido à complexidade do problema, as métricas associadas à classificação são importantes para se mensurar a qualidade dos resultados. Porém, individualmente, as métricas são insuficientes para uma análise mais detalhada. Por exemplo, a ferramenta TribeMCL obteve uma maior quantidade de verdadeiro positivos, mas isso não necessariamente implica no melhor agrupamento para o problema. Um caso hipotético extremo, caso um algoritmo classificasse todas as sequências como homólogas o total de VP seria máximo, contudo o total de falso positivos seria extremamente alto.

A classificação considerando o conjunto completo de sequências (hipotéticas e não hipotéticas) obteve os piores resultados. Isto ocorreu possivelmente devido à presenta de famílias constituídas por anotação mistas, que tenham sequências anotadas com funções conhecidas e desconhecidas (hipotéticas). Isso pode indicar um possível erro de classificação dos algoritmos, ou apenas que essas sequências de fato são homólogas e que o problema em questão está no processo de anotação das sequências. Essa mistura específica de anotações não é muito comum, mas foram encontrados alguns casos relevantes envolvendo proteínas de fagos no conjunto de *S. pyogenes*, já os potenciais problemas de agrupamento no conjunto das *Xanthomonadaceae* estão mais relacionados a elementos móveis e proteínas relacionadas com a composição da membrana plasmática.

Com exceção das métricas de sensibilidade e eficiência, o Multilayer Clustering obteve resultados melhores ou no mínimo equivalentes aos apresentados pelo TribeMCL, ao se considerar o primeiro nível de agrupamento realizado pelo algoritmo proposto. As camadas de corte se mostraram úteis para o processo de agrupamento, uma vez

que os resultados demonstram que a progressão de camadas melhora consideravelmente praticamente todas as métricas, porém aumentar ainda mais o número de camadas poderia resultar em possíveis pioras nas métricas (por isso foi necessária a especificação de um método criterioso para a realização dos cortes, conforme já apresentado). Ao se considerar o tratamento de sequências multidomínio realizando uma subdivisão de alguns agrupamentos produzidos, foi possível observar uma melhorara significativamente nas métricas de avaliação do Multilayer Clustering após a separação dos domínios, indicando a importância desta etapa para a melhor separação das famílias.

Por razões já discutidas anteriormente, como a falta de uma curadoria para a anotação das sequências, esses resultados não conseguem precisar o desempenho de ambos algoritmos. Contudo, esse experimento possibilita entender o comportamento dos algoritmos no subespaço dos dados avaliado. Adicionalmente, a abordagem apresentada preserva a estrutura do grafo permitindo outras análises topológicas, por exemplo, a identificação de sequências multidomínios e seus respectivos domínios. A identificação de domínios mostrou potencial para aprimorar a identificação de grupos. Destaca-se que mesmo sem nenhuma estratégia definida para a especificidade do caso discutido, a ferramenta TribeMCL também apresentou resultados muito bons.

4.2 Desempenho da execução do pipeline

O GTACG fornece resultados bastante completos, quando comparado com outros arcabouços (PAGE et al., 2015; CHAUDHARI; GUPTA; DUTTA, 2016; ZHAO et al., 2014), abrangendo diferentes fases de uma pesquisa focada em pan-genomas. Em geral, esse processo tem início com a reanotação automática das sequências, sucedida pela busca dos alinhamentos locais. Essas etapas são as mais custosas do ponto de vista computacional.

O tempo necessário para a anotação automática, assim como a qualidade e especificidade dos resultados, é dependente da ferramenta escolhida pelo usuário. Essa etapa é bastante custosa computacionalmente e, dependendo da ferramenta escolhida, pode exigir um esforço manual considerável do usuário. Entretanto, esta é uma etapa inevitável para minimizar erros metodológicos em muitos pipelines de ferramentas baseadas na identificação de genes homólogos.

Para medir o desempenho computacional das etapas subsequentes, foram preparados testes com cinco conjuntos de genomas do gênero *Xanthomonas* escolhidos com um total de 10, 20, 30, 40 e 50 genomas. A descrição destes genomas está presente no material Anexo A, seção A.3. O computador escolhido para a execução foi um *Intel*(R) *Xeon*(R) E5-2620 com 24 núcleos e 64GB de memória RAM e, para as etapas que permitem execução de forma concorrente, foram mensurados os tempos utilizando 5, 10, 15 e 20 núcleos. Os tempos de execução são apresentados na tabela 5.

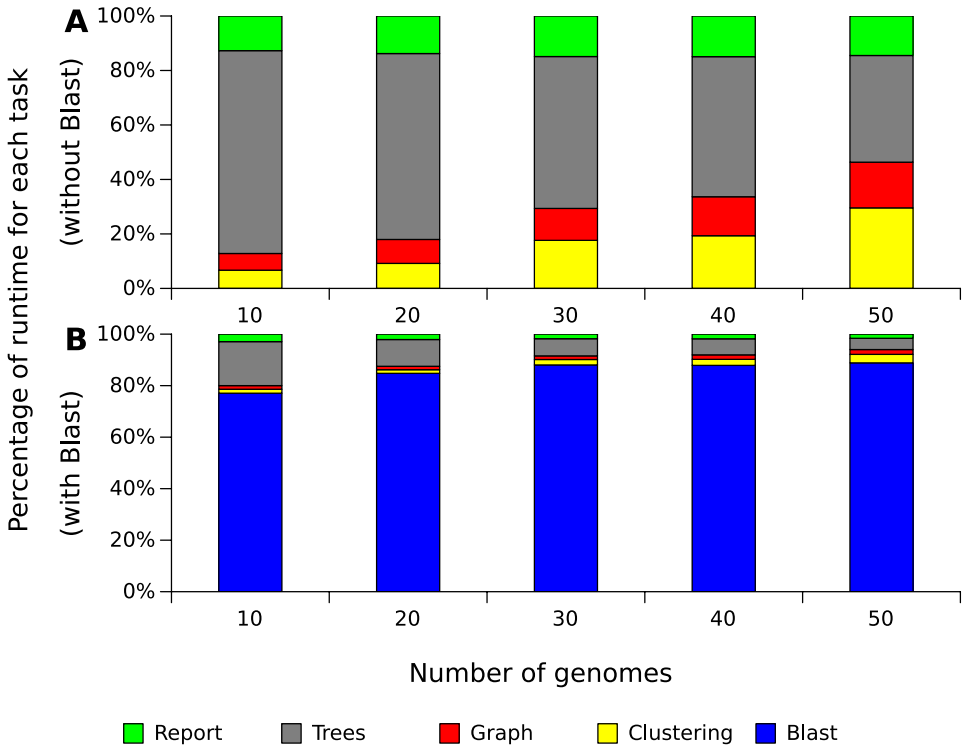
Tabela 5 – Tempo de execução para os experimentos sintéticos com 10, 20, 30, 40 e 50 genomas. Todas as execuções foram feitas em um computador com processador *Intel*(R) *Xeon*(R) E5-2620 com 24 núcleos. Os tempos resultantes estão apresentados na forma de segundos.

Etapa	Número de núcleos	Genomas 10	20	30	40	50
		10	20	30	40	
Listar sequências	–	4,04	6,87	9,68	12,59	16,14
Busca com BLAST	5	2497,39	10183,3	23178,64	42025,58	66661,24
	10	1289,03	5110,52	11751,47	21296,11	33769,58
	15	1023,38	3921,07	9419,43	16392,82	26008,75
	20	885,14	3703,94	8775,91	15648,00	25039,39
Busca com MMseqs2	5	427,51	1054,93	1875,99	2863,77	4010,84
	10	221,43	548,52	969,83	1480,42	2070,17
	715	180,64	445,35	786,23	1203,90	1678,29
	20	171,60	419,92	742,24	1131,78	1591,18
Clusterização das sequências	1	31,05	190,48	903,73	2349,09	5555,40
	5	17,97	77,23	288,24	674,20	1567,47
	10	17,14	68,39	236,13	511,04	1140,76
	15	16,68	63,22	207,29	438,43	977,97
	20	17,63	61,04	210,34	417,69	929,38
Exportar clusters como grafo	–	16,13	58,41	139,76	309,14	527,36
Produção de alinhamentos e inferência das filogenias	1	2030,26	5287,93	7399,21	11281,53	12476,08
	5	429,53	1127,85	1576,07	2425,43	2674,81
	10	250,71	617,00	848,27	1413,39	1536,44
	15	204,48	498,76	750,32	1214,65	1321,51
	20	196,08	453,62	664,70	1110,95	1231,06
Consolidação de todos os resultados em um website estático	1	79,04	221,93	355,99	818,63	926,91
	5	36,63	102,83	194,98	402,33	476,25
	10	37,09	99,71	186,99	374,31	460,37
	15	35,94	95,37	178,98	361,77	449,08
	20	33,50	91,64	177,46	322,47	455,82

Fonte: Santiago et al. (2019)

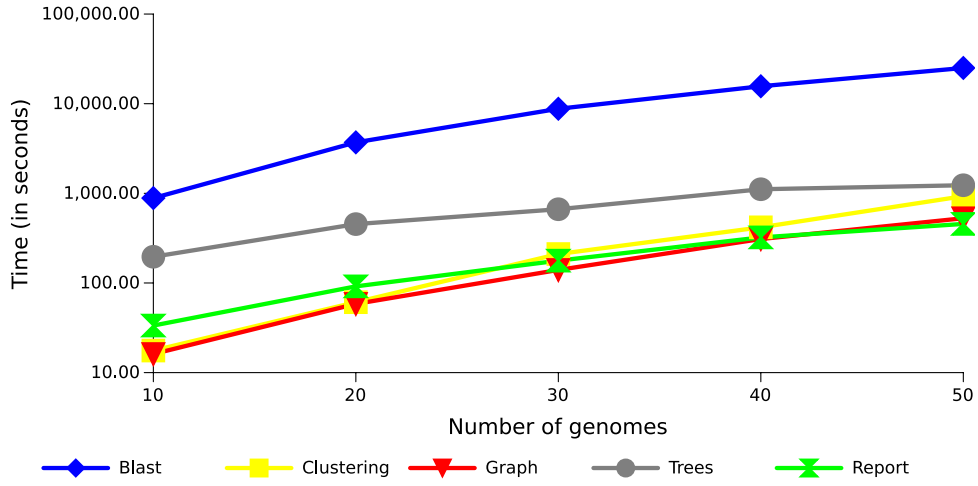
A produção dos alinhamentos locais (de todas as sequências de aminoácidos contra todas estas sequências) foi realizada utilizando BLAST (*blastp*). Esta etapa corresponde ao maior custo computacional de tempo de todo o pipeline, consumindo entre 75% e 95% do tempo de execução para estes conjuntos de dados (como visto na figura 19 para as execuções de 20 núcleos). Embora esse processo possa ser acelerado por meio de execuções paralelas, a tendência desse consumo é quadrática (Figura 20), justificável pelo fato do número de alinhamentos também ter um crescimento quadrático em relação ao crescimento linear no número de genomas. Uma alternativa viável ao BLAST é o MMseqs2 (STEINEGGER; SÖDING, 2017) com sensibilidade definida em 7,5. O MMseqs2 consome consideravelmente menos tempo que o BLAST (chegando a ser de 30 a 35 vezes mais rápido) e os resultados produzidos se mantêm com qualidade similar, não prejudicando as conclusões obtidas nesta avaliação de desempenho e nos estudos de caso discutidos anteriormente.

Figura 19 – Tempo de execução do GTACG relativo às principais etapas considerando conjuntos com diferentes quantidades de genomas de *Xanthomonas*. Esses resultados foram obtidos usando um computador com processador *Intel*(R) *Xeon*(R) E5-2620. Este computador tem 24 núcleos, mas estes resultados foram produzidos utilizando 20 núcleos. Os resultados estão separados em duas seções, na seção (A) estão os tempos de execução desconsiderando a etapa de execução do BLAST, já na seção (B) está incluso o tempo de execução do BLAST (que é a maior parte do tempo consumido).



Fonte: Santiago et al. (2019)

Figura 20 – Tempo de execução do GTACG relativo às principais etapas considerando conjuntos com diferentes quantidades de genomas de *Xanthomonas*. Apresentação dos resultados como uma curva de crescimento em função do tamanho do conjunto de genomas.



Fonte: Santiago et al. (2019)

As demais etapas também mostram uma tendência no máximo quadrática de consumo de tempo em relação ao número de genomas. Com exceção da produção de alinhamentos locais, a etapa de maior custo computacional é a preparação de todos os alinhamentos múltiplos (um para cada família) e das filogenias de cada uma das famílias, porém essa etapa tem uma tendência mais linear do que as anteriores. As árvores filogenéticas foram inferidas utilizando o FastTree, uma ferramenta com foco em baixo consumo de tempo (mesmo em casos de muitas sequências ou sequências longas). Outra ferramenta que pode ser utilizada para esta etapa é a PhyML, porém a execução desta ferramenta leva bem mais tempo do que a execução do FastTree.

Embora o GTACG gaste mais tempo para ser executado quando comparado com outros arcabouços mais simples, como o Roary (PAGE et al., 2015), o BPGA (CHAUDHARI; GUPTA; DUTTA, 2016) ou o PanGP (ZHAO et al., 2014), o GTACG fornece mais informações para os usuários, sem que para isso seja necessário complementar a análise com outras ferramentas externas, além de diferentes resultados e ferramentas para análise de pan-genomas de forma simples e prática para usuários não programadores.

4.3 Comparação entre ferramentas de análise de pan-genomas

O estudo de pan-genomas data de mais de uma década (VERNIKOS et al., 2015). No decorrer destes anos de pesquisa na área, alguns arcabouços computacionais foram desenvolvidos para análise de pan-genomas, com abordagens também baseadas em grupos de famílias homólogas (ou ortólogas). Contudo, muitos desses trabalhos estão limitados a métricas estatísticas globais, como diferentes formas de categorização do core-genoma ou a contagem de genes únicos dentro do conjunto analisado (PAGE et al., 2015; ZHAO et al., 2014; LAING et al., 2010; BENEDICT et al., 2014). Outra abordagem comum a essas ferramentas é a busca por uma filogenia dos dados de entrada, considerando diferentes técnicas e não se limitando ao uso de alguns marcadores filogenéticos (CLARRIDGE, 2004).

Famílias de sequências ou genes homólogos têm uma ampla gama de informações a serem mineradas, que não se restringem ao core-genoma. Neste contexto se torna mais importante a disponibilidade de mecanismos de busca sofisticados e que consigam encontrar informações valiosas sobre famílias de genes acessórios. Apesar de muitos arcabouços terem sido desenvolvidos no decorrer dos anos, a mineração de dados sobre as famílias é uma limitação em boa parte desses arcabouços. Alguns trabalhos, apesar de discutirem problemas similares, utilizam métodos manuais (ILINA et al., 2013; PRASANNA; MEHRA, 2013; VLIET, 2017), talvez pela falta de uma metodologia já estabelecida para auxiliar o trabalho do pesquisador.

Ao se considerar os métodos automáticos para análise de pan-genomas e buscas sobre famílias de genes (alguns deles listados no quadro 4.3), se destacam o PGAT (ZHAO et al., 2018), o PanX (DING; BAUMDICKER; NEHER, 2018), e o método de Obolski et al. (2018). Embora o PGAT forneça uma ampla gama de buscas por genes com base em características específicas, ela é limitada, podendo ser feita apenas em um conjunto específico de genomas. Uma das principais limitações do PGAT reside na rigidez do mecanismo de busca, não permitindo a busca por resultados aproximados, assim como a burocracia em testar buscas com objetivos diferentes. Limitações que também são compartilhadas pelo BPGA (CHAUDHARI; GUPTA; DUTTA, 2016) que realiza buscas por características fenotípicas, mas de forma rígida. Por exemplo, caso algum fenótipo não tenha sido anotado corretamente (ou não tenha sido observada sua expressão) pelo

usuário, ele não será facilmente encontrado, por exigir muitas buscas consecutivas para resolver esse problema. Embora o PGAT apresente seus resultados na forma de um *website*, as especificidades dos resultados (como resultados de uma busca) não são facilmente compartilhados. O PanX também apresenta seus resultados na forma de um *website*, mas este é interativamente mais dinâmico do que o do PGAT. Porém, os parâmetros que norteiam as buscas no PanX são limitados a estatísticas sobre as famílias, como número de genomas. Uma vantagem do PanX é a visualização das filogenias das famílias personalizadas com base nas anotações fenotípicas. Por fim, o método de Obolski et al. (2018) utiliza o algoritmo *Random Forest* (BREIMAN, 2001) para encontrar famílias mais relacionadas com a invasividade anotada para um conjunto de genomas de *Streptococcus pneumoniae*.

Quadro 1 – Comparação das principais funcionalidades de alguns arcabouços computacionais para estudo genômicos.

Funcionalidades	Arcabouços computacionais								
	GTACG	BPGA	PanX	PGAT	PanGP	PGAP	Panseq	ITEP	get_homologues
Identificação de genes específicos a fenótipos – lista	X	X		X					X
Identificação de genes específicos a fenótipos – métricas	X								
Distribuição do core-genoma, genes únicos e acessórios	X	X	X						
Análise do perfil do pan-genoma	X	X			X	X			X
Tamanho do core e do pan-genoma	X	X	X			X	X		X
Extração do core-genoma, genes únicos e acessórios	X	X						X	
Análise filogenética	X	X	X			X	X	X	X
Clusterização de genes	X	X	X	X		X	X	X	X
Detalhamento com diferentes níveis dos genes	X		X	X				X	
Dados de entrada fornecidos pelo usuário	X	X			X	X	X		X
Facilidade de compartilhar resultados	X			X			X		
Integração com códigos do Roary	X								
Preparação dos dados	C	C	C	N	G	C	C	C	C
Interface do usuário	W	G	W	W	G	G	G	G	G

Preparação dos dados: C – Linha de comando; G – Interface gráfica.
Interface do usuário: N – Não aplicável; W – Website; G – Saída gráfica.

Fonte: Santiago et al. (2019)

O PanSeq (LAING et al., 2010), assim como o PanX e o PGAT, também disponibiliza os resultados de forma fácil (por meio de URLs), mas como um serviço para se obter resultados de forma limitada a arquivos com resultados pontuais, sem customização e sem interação com o usuário. De forma geral, os demais arcabouços disponíveis são bastante focados em uma experiência limitada a comandos de texto, como o ITEP (BENEDICT et al., 2014) ou get_homologues (CONTRERAS-MOREIRA; VINUESA, 2013), ou interfaces pouco interativas, como o PGAP (ZHAO et al., 2012) que foi recentemente estendido para interfaces gráficas (ZHAO et al., 2018).

Baseada na descrição das qualidades e limitações das ferramentas mencionadas anteriormente (e listadas no quadro 4.3), o GTACG combina vários benefícios de todas elas, além de propor um algoritmo de agrupamento de sequências dedicado ao problema. Além disso, o GTACG se destaca por facilitar a visualização de dados e o compartilhamento de resultados de pesquisa. Embora não seja possível cobrir toda a diversidade de ferramentas destinadas ao estudo de pan-genoma, procura-se contornar essa limitação estruturando o desenvolvimento em um ambiente aberto e facilmente modificável, exigindo menos esforço para programar novos conteúdos, reduzindo assim as dificuldades impostas por algumas ferramentas voltadas ao estudo da biologia de sistemas (HILLMER, 2015). O código fonte do GTACG, bem como documentação e arquivos complementares estão disponíveis na Internet[1,2]

4.4 Análise dos estudos de caso

Devido à estratégia adotada pelo arcabouço desenvolvido de agrupar os genes em famílias e permitir diferentes análises com base em informações de cada genoma (como características fenotípicas compartilhadas por alguns dos genomas analisados), o arcabouço desenvolvido permite que qualquer busca focada nestas informações sejam feitas de forma simples e eficiente, facilitando a descoberta de conhecimento sobre possíveis mecanismos genéticos associados aos fenótipos.

Considerando o estudo de caso com 161 genomas da família *Xanthomonadaceae*, os resultados obtidos em si não são suficientes para concluir a participação de qualquer uma das famílias na expressão do fenótipo, mas são um ponto de partida para guiar novos

[1] Back-end do GTACG: ⟨https://github.com/caiorns/GTACG-backend⟩.
[2] Front-end do GTACG ⟨https://github.com/caiorns/GTACG-frontend⟩.

estudos laboratoriais. A importância em si desses resultados está justamente em diminuir o escopo de uma pesquisa de milhares de genes, para apenas algumas dezenas deles.

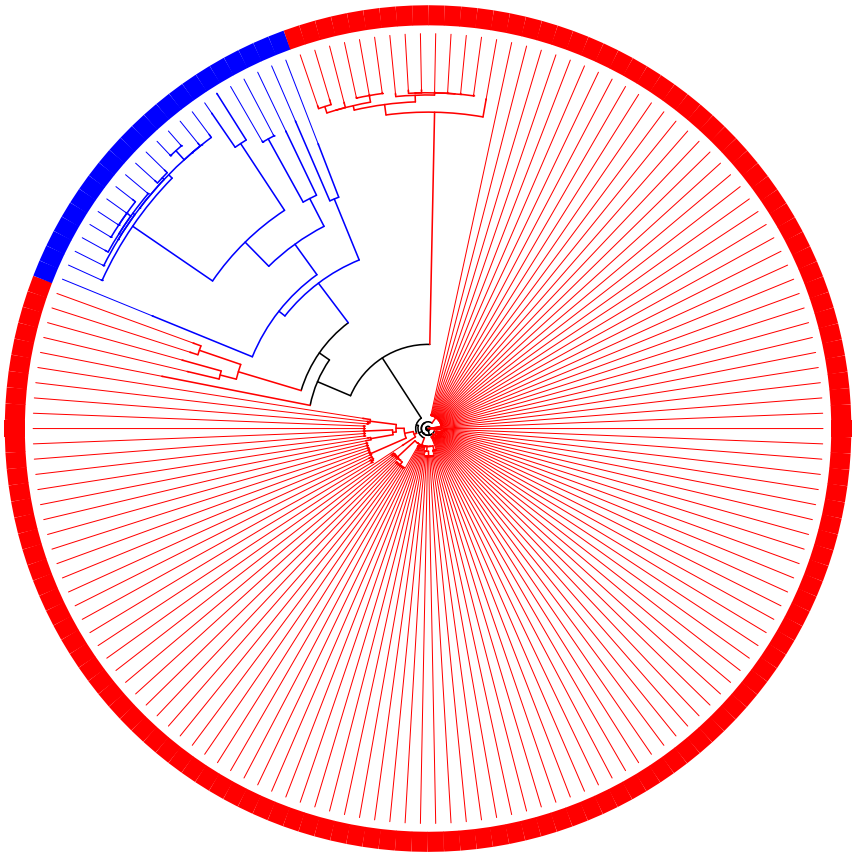
O mesmo comportamento visto nas filogenias é refletido na composição das famílias. Mesmo que os dois grupos (associados e não associados a plantas) estejam, em geral, bem divididos, alguns ramos específicos são compostos por genomas dos dois grupos. Foram identificadas 19 famílias que são únicas aos genomas associados a plantas e estão presentes em pelo menos 90% deles. *X. mangiferaeindicae* é o único a não ter um gene em nenhuma destas famílias, sendo a única exceção a 15 das 19 famílias de genes de genomas associados a plantas. Das quatro famílias restantes, uma não contém apenas a *X. albilineans*, um microrganismo amplamente estudado e cuja ausência nessa família pode ser resultado de um processo evolutivo baseado em redução de genoma (PIERETTI et al., 2009). Em duas outras famílias, os genomas presentes são os mesmos descritos pela *supertree* que não estavam em ramos mistos (agrupados em conjunto com genomas não associados a plantas). Considerando essas 19 famílias de genes, elas podem ser importantes para a interação metabólica com plantas e, portanto, *X. mangiferaeindicae* pode ter se adaptado para utilizar um via metabólica alternativa, assim como a *X. albilineans* pode ter se adaptado a utilizar um subconjunto de genes reduzido dessas famílias. Por fim, uma dessas famílias não contém nenhuma das quatro cepas de *X. fragariae* (além de não conter *X. mangiferaeindicae*).

Destacam-se também as famílias que contêm todos os genomas associados a plantas, mas que também incluem poucos genomas não associados a plantas. Existe uma família que contém todos os três genomas não associados a plantas que foram aproximados a genomas associados a plantas pelo método da *supertree*: *P. suwonensis* 11-1, *P. suwonensis* cepa J1, e *P. spadix* BD-a59. Também existem outras oito famílias que adicionam, aos já citados três genomas anteriores, genes de *S. nitritireducens*, *Stenotrophomonas sp.* KCTC 12332 e *S. acidaminiphila*. Como o número de genomas não associados a plantas são minoritários nestas famílias, isso leva a hipótese de que essas famílias podem ser importantes para permitir a associação com plantas, mas talvez alguns desses microrganismos não expressem esses genes, ou esses genes podem pertencer a uma via metabólica que depende de outros genes ausentes em alguns desses genomas e, portanto, esses organismos não apresentariam um dado fenótipo.

Baseado nos alinhamentos múltiplos produzidos para cada família, destacam-se nove casos com mutações nos aminoácidos dos genes associados a plantas com dissimilaridade

maior que 1%, indicando possíveis pequenos trechos de recombinações ou mutações pontuais, como polimorfismos de nucleotídeo único (em inglês *Single-Nucleotide Polymorphism* – SNP). Por outro lado, dissimilaridades menores que o limiar 1% não são muito conclusivas, resultando em muitas mutações não exclusivas. Adicionalmente, foram encontradas 13 famílias em que suas filogenias dividem perfeitamente em dois grupos todas as suas sequências de acordo com os grupos anotados, como exemplificado pela figura 21, que apresenta uma dessas filogenias. As filogenias obtidas por estas 13 famílias não diminuem necessariamente a confiança das filogenias dos genomas encontradas pelos outros métodos de inferência, mas como a filogenia foi extraída a partir dos aminoácidos, esses resultados indicam uma diferença significativa nos aminoácidos destas famílias que não foi percebida pela métrica de dissimilaridade, pois se trata de uma diferençam combinada de aminoácidos.

Figura 21 – Filogenia de uma família de genes ortólogos do conjunto de 161 genomas de *Xanthomonadaceae*. Nesta família os genes pertencentes aos genomas associados a plantas são agrupados em um único ramo, de forma isolada dos genes dos genomas não associados a plantas. As proteínas, neste caso, foram todas anotadas como "N(6)-L-threonylcarbamoyladenine synthase".



Fonte: Santiago et al. (2019)

Já com relação os conjuntos de genomas de *S. pyogenes* e suas múltiplas características de interesse para esse estudo, os resultados mais conclusivos foram os relacionados ao genótipo *emm*. Os três métodos utilizados para inferir a filogenia conseguiram isolar cada um dos tipos de genótipos *emm* em ramos únicos, em acordo com o que já era esperado na literatura (HOLLINGSHEAD et al., 1994). Este genótipo está em uma região grande (de 3 a 6 mil pares de bases) e de alta diversidade, e isso dificulta que o processo de identificação de famílias de genes consiga delimitar essa região como uma família única e também afeta de forma similar o processo de anotação automática de genes. Entretanto, 15 famílias ortólogas conseguiram encontrar filogenias equivalentes. Encontrar essas 15 famílias de forma fácil é importante para definir uma forma mais simples (mais regiões mais curtas e de menor diversidade) que podem agir, neste e outros conjuntos, como marcadores filogenéticos.

O mesmo processo não se reflete de forma tão simplificada no estudo das doenças causadas pelo *S. pyogenes*, muito em parte por esse fenótipo não ter, necessariamente, relações diretas com a filogenia. Algumas hipóteses podem ser relevantes para melhor entender esse processo, considerando, é claro, que o fenótipo tenha sido corretamente anotado. Pelo fato do fenótipo expressar uma relação dos microrganismos com o sistema imunológico do hospedeiro, existem diversos fatores que podem influenciar na infecção (MENENDEZ; FINLAY, 2007; DOBRINDT et al., 2004; NOVERR; HUFFNAGLE, 2004), como fatores ambientais e da biologia do hospedeiro. Outro fator a ser considerado é que a anotação com relação a doenças é bastante vaga, pois, uma vez que uma bactéria é sequenciada de um tecido em particular, isso prova apenas que essa bactéria é de fato capaz de infectar esse tecido, porém esta informação nada diz a respeito sobre a infecção de outros tecidos. Por exemplo, uma bactéria sequenciada a partir de uma meningite, não traz informação sobre a capacidade desta mesma bactéria causar endometrite caso a exposição do hospedeiro a bactéria se desse por uma via diferente (ou caso o hospedeiro tivesse uma imunossupressão de alguma natureza). Por esse motivo, existem muitas lacunas de informações que não foram anotadas o que torna prematuro interpretar esses grupos de genomas como mutuamente exclusivos.

Por causa dos motivos discutidos, a descoberta de informações sobre as doenças é bastante problemática, e isso se reflete nos resultados encontrados de famílias compartilhadas exclusivamente pelos genomas que causam as doenças. Os dados mais expressivos são obtidos de grupos unitários ou formados por poucos genomas e, portanto, são estatistica-

mente pouco confiáveis. Uma forma de identificar dados relevantes sobre esses grupos de genomas maiores e pouco informativos, é procurar pelas famílias que são compartilhadas por todos os genomas de um determinado grupo, mas que possuem um número reduzido de outros genomas envolvidos. Outra abordagem interessante é encontrar poucas famílias que quando combinadas são exclusivas apenas aos genomas de um grupo fenótipo. A doença GlomeruloNefrite Pós-Estreptocócica Aguda, por exemplo, foi anotada como causada por três diferentes cepas e não foi encontrada nenhuma família que fosse exclusiva destes três genomas. Além disso, a menor família que abarca todos esses três genomas possui outros 14 genomas não pertencentes ao grupo. Entretanto, pelo método de combinação de famílias foram encontradas 19 possíveis combinações de famílias que quando juntas são exclusivas a esses três genomas, sendo que em todas as combinações existem três principais famílias: uma mesma proteína hipotética, na maioria das vezes, combinada com outras proteínas de fagos de endopeptidase ou hialuronidase, e por fim combinadas com outras famílias variadas. Essa combinação de famílias pode ser um importante fator de virulência dessas cepas para explicar esse fenótipo em específico ou outros fenótipos associados a doenças.

Utilizando o mesmo princípio de encontrar famílias que mais se aproximam, as famílias relacionadas à febre reumática aguda também se mostram relacionadas à faringite (sete famílias compartilhadas) seguidas de algumas poucas famílias relacionadas à bacteremia, fascite necrosante e muitos outros grupos. Outro resultado interessante diz respeito à síndrome do choque tóxico estreptocócico, em que sete famílias apresentaram a mesma configuração com genomas associados a infecção do cérebro fluido espinhal, fascite necrosante, escarlatina, faringite e dermatite, sendo (parte de) uma possível via de todas essas doenças.

Os grupos com maior número de genomas, como faringite e fascite necrosante, são bastante disseminados entre famílias de outros grupos de doenças. Isso pode indicar que os fatores que levam a uma ou outra infecção podem ser causados por fatores externos à composição das famílias, ou até mesmo ser causados por vias metabólicas complexas, difíceis de serem identificadas.

4.5 Descrição funcional das proteínas encontradas exclusivamente em genomas de Xanthomonadaceae *associados a plantas*

Entre as 19 proteínas identificadas em pelo menos 134 dos 139 genomas dos microrganismos associados a plantas dos 161 tratados nos estudos de casos, oito destas famílias estão envolvidas na degradação de N-glicanos. Curiosamente, os genes ligados à degradação de N-glicanos se encontram na mesma região genômica, constituindo um agrupamento (*nix*) em conjunto com o *cutC* (resistentes a cobre) e são responsáveis pela clivagem dos N-glicanos em diferentes ligações glicosídicas (Quadro 2 e Figura 22). A interação de patógenos de plantas é propiciada pela evolução das proteínas ligadas à virulência bacteriana para induzir a virulência e modular a resposta imune das plantas, isso concomitante com a evolução das proteínas vegetais para reconhecer os efeitos da infecção bacteriana e induzir resposta imunológica especializada levando à resistência. Os receptores de reconhecimento de padrões (em inglês *pattern-recognition receptors* – PRR) são responsáveis por reconhecer padrões moleculares associados a patógenos (em inglês *Pathogen-associated molecular pattern* – PAMP) e pela ativação de gatilhos imunológicos (em inglês *pathogen-triggered immunity* – PTI). Häweker et al. (2010) demonstraram que os PRR precisam de N-glicosilação para mediar a imunidade da planta. Pela degradação de N-glicanos associados aos receptores das plantas, o hospedeiro perde a capacidade de reconhecer a infecção e ativar a resposta imune, permitindo assim um maior sucesso na colonização e adaptação dessas bactérias no hospedeiro.

Adicionalmente, outras proteínas encontradas estão envolvidas na adaptação, incluindo duas peptidases (homólogas a XAC0609 (ZHOU; YAN; WANG, 2017b) e a PepQ-XAC2545) e três proteínas hipotéticas (homólogas a XAC2544, XAC4076 e XAC4164, presentes no quadro 2). Análises da sequência relacionada com a XAC0501 revelaram que este gene codificado pelas *LesA/LipA* é um fator de virulência chave necessário para a patogenicidade de *Xylella fastidiosa* em videiras (NASCIMENTO et al., 2016), ou de *Xanthomonas citri* em frutos cítricos (ASSIS et al., 2017) e de *Xanthomonas oryzae* em arroz (APARNA et al., 2009). Outros quatro genes também podem estar relacionados com a adaptação. O *hspA* tem sido descrito tanto como um *chaperone* muito importante quanto como um agente de proteção durante o armazenamento de proteínas na *Xanthomonas campestris* (LIN et al., 2010). O *cyoD* codificado por um *citocromo O ubiquinol oxidase subunidade IV*, que é um componente da cadeia respiratória aeróbica que é predominante

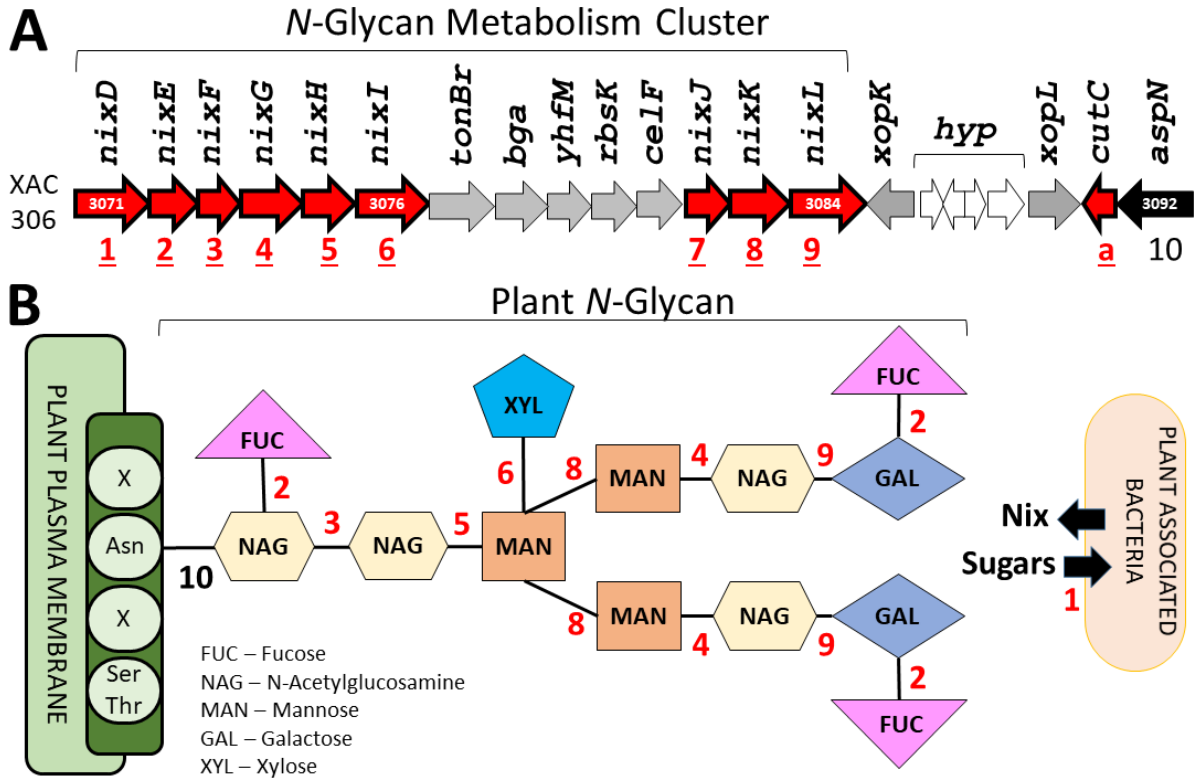
Quadro 2 – Caracterização das 18 famílias de proteínas identificadas como exclusivas aos genomas de bactérias associados a plantas, considerando o estudo de caso dos 161 genomas de *Xanthomonadaceae*.

Função	Gene	Ref. Locus Tag	# Genomas	# Parálogos	Via Metabólica	PS	Referências
Conserved hypothetical protein (putative lipase)	*lesA* (*lipA*)	XAC0501	134	27	Lipid metabolism	N	(NASCIMENTO et al., 2016), (ASSIS et al., 2017), (APARNA et al., 2009)
Peptidase M16 family / Zinc protease / Insulinase family protein	—	XAC0609	138	1	Peptidases	S	(ZHOU; YAN; WANG, 2017a)
Low molecular weight heat shock protein / Molecular chaperone	*hspA*	XAC1151	138	1	Chaperones and folding catalysis	N	(LIN et al., 2010)
Cytochrome O ubiquinol oxidase subunit IV	*cyoD*	XAC1261	138	2	Oxidative phosphoryla-tion	N	(LUNAK; NOEL, 2015)
Conserved hypothetical protein	—	XAC2544	137	2	Unknown function	S	—
Predicted 4-hydroxyproline dipeptidase / Xaa-Pro aminopeptidase	*pepQ*	XAC2545	138	1	Metallo peptidases	N	—
Alpha-L-fucosidase	*nixE*	XAC3072	138	1	N-glycan metabolism	S	(ASSIS et al., 2017), (DUPOIRON et al., 2015), (BOULANGER et al., 2014)
Hypothetical protein (putative glycosyl-hydrolase)	*nixF*	XAC3073	138	1	N-glycan metabolism	S	(ASSIS et al., 2017), (DUPOIRON et al., 2015), (BOULANGER et al., 2014)
Beta-hexosaminidase / Beta-N-acetylglucosaminidase	*nixG*	XAC3074	138	1	N-glycan metabolism	S	(DUPOIRON et al., 2015), (BOULANGER et al., 2014)
Beta-mannosidase	*nixH*	XAC3075	138	3	N-glycan metabolism	S	(DUPOIRON et al., 2015), (BOULANGER et al., 2014)
Beta-glucosidase-related glycosidases / Gluca-beta-glucosidase	*nixI*	XAC3076	138	2	N-glycan metabolism	S	(ASSIS et al., 2017), (DUPOIRON et al., 2015), (BOULANGER et al., 2014)
Hypothetical protein (putative glycosyl-hydrolase)	*nixJ*	XAC3082	138	4	N-glycan metabolism	S	(DUPOIRON et al., 2015), (BOULANGER et al., 2014)
Alpha-1,2-mannosidase	*nixK*	XAC3083	138	1	N-glycan metabolism	N	(DUPOIRON et al., 2015), (BOULANGER et al., 2014)
Beta-galactosidase	*nixL*	XAC3084	138	1	N-glycan metabolism	N	(ASSIS et al., 2017), (DUPOIRON et al., 2015), (BOULANGER et al., 2014)
Cytoplasmic copper homeostasis protein CutC	*cutC*	XAC3091	138	2	Copper metabolism	N	—
3-isopropylmalate dehydrogenase / Isocitrate dehydrogenase	*leuB*	XAC3456	134	1	Leucine biosynthesis	N	(MOREIRA et al., 2017), (LAIA et al., 2009)
Integral membrane protein	—	XAC4076	134	1	Unknown function	N	—
N-acetylglucosamine-regulated / TonB-dependent receptor	*nixD*	XAC4131/ 3071	138	10	TonB receptors/ N-glycan metabolism	S	(BLANVILLAIN et al., 2007)
Conserved hypothetical protein	—	XAC4164	137	1	Unknown function	S	(JALAN, 2012)

PS – Peptídeo sinal; S – Sim, N – Não.

Fonte: Santiago et al. (2019)

Figura 22 – Identificação de genes relacionados a degradação de N-glicanos. (A) Agrupamento de genes metabólicos de N-glicanos no genoma Xac306. Em vermelho estão os genes identificados como exclusivos aos genomas associados a plantas. Os números de 1 a 10 identificam todos os genes relacionados a degradação de N-glicanos. (B) Modelo estrutural dos N-glicanos de plantas. Os números de 1 a 10 identificam pontos catalíticos das proteínas codificadas pelos genes descritos em A. Asn – Resíduo de asparagina. Ser/Thr – Resíduo de Serina e Treonina. X – Outros resíduos.



Fonte: Santiago et al. (2019)

quando células crescem em alta aeração (LUNAK; NOEL, 2015). O *leuB* codificado por uma 3-isopropilmalato desidrogenase que foi super regulada (*upregulated*) em *Xanthomonas axonopodis* pv. *citri* (Xac) 1, 3 e 5 dias depois da inoculação (MOREIRA et al., 2017). Quando mutada, a ausência de *leuB* mostrou redução da virulência de Xac no hospedeiro compatível (LAIA et al., 2009). Apenas homólogos à XAC4076 codificados por uma proteína completa da membrana não foram investigados em outros estudos.

Por fim, a última famílias proteicas exclusiva aos genomas associados a plantas é codificada por um receptor *TonB-dependent* (em inglês *TonB-dependent receptor* – TBDR) homólogo ao XAC4131. Blanvillain et al. (2007) predisseram 72 TBDR na *Xanthomonas campestris*, vários deles pertencentes a lócus de utilização de carboidratos de plantas como a sacarose, compostos de parede celular vegetal e pectina. Assim, as bactérias também podem

usar os subprodutos como fonte de energia por meio da internalização dos monômeros através de TBDR, uma proteína de membrana externa conhecida principalmente pelo transporte ativo de moléculas. Destaca-se que dez parálogos deste gene foram encontradas nos genomas investigados (Quadro 2). Um desses parálogos é codificado pelo gene XAC3071 no genoma Xac306, o que corresponde ao *nixD*, o primeiro gene do agrupamento descrito anteriormente (Figura 22A). É possível que esses genes TBDR estejam envolvidos na internalização de açúcares derivados da degradação de N-glicanos, que poderiam servir como uma fonte alternativa de carbono após a supressão imune da planta.

Esta análise do repertório dos genes identificados pelo GTACG permite inferir que o arcabouço computacional desenvolvido se mostrou eficiente na busca de informações genéticas correlacionadas com informações fenotípicas, uma vez que os genes identificados como exclusivos a genomas associados a plantas já foram descritos como capazes de modular a adaptação bacteriana à planta hospedeira.

5 Conclusão

No decorrer deste texto foi apresentado o GTACG (*Gene Tags Assessment by Comparative Genomics*) um arcabouço computacional que contempla todo o ciclo de vida, do ponto de vista computacional, de uma pesquisa sobre genômica comparativa de bactérias. O principal foco que norteou todo o processo de desenvolvimento deste arcabouço foi que a partir de genomas distribuídos em um mesmo ramo evolutivo, o pesquisador conseguisse encontrar características genéticas relacionadas com características fenotípicas. Para isso, foi definida uma ampla gama de métricas e ferramentas de buscas que permitem ao pesquisador extrair informações acerca de todo o pan-genoma.

O projeto foi baseado em uma abordagem de famílias de genes, que são construídas a partir de alinhamentos locais. Portanto, esses genomas devem ter sido anotados (preferivelmente de forma automática para evitar problemas metodológicos) e em seguida suas CDS devem ser alinhadas. A partir deste ponto o projeto possui três etapas bem definidas. A primeira delas é o agrupamento de sequências, que tem como objetivo a identificação das famílias de genes. Foi desenvolvido um algoritmo próprio para o agrupamento de sequências, com diferencial de assumir que as sequências são provenientes de genomas distribuídas sobre um mesmo ramo evolutivo e, por esse motivo, espera-se que os dados de distribuam de forma mais homogênea e densa. Esta primeira etapa é fundamental, pois dela se derivam os resultados que guiam as análises. O algoritmo desenvolvido se mostrou equivalente ao TribeMCL, um algoritmo bem estabelecido para o agrupamento de sequências e, em certas condições, foi capaz de superá-lo, com a vantagem de ter uma estratégia explícita para lidar com proteínas multidomínio.

A etapa seguinte é a comparação dos genomas. Nesta etapa os genomas são processados com base nas famílias de genes a fim de gerar resultados comparativos, como suas filogenias. As famílias de genes são analisadas e são estabelecidas métricas, tanto sobre estatísticas básicas (como número de genomas ou comprimento das sequências) quanto métricas de correlação sobre os grupos fenótipos estabelecidos pelo usuário.

Por fim, a terceira etapa é destinada à visualização dos resultados. Esta etapa foi construída com base em um *website* estático e interativo. Essa abordagem possui uma série de benefícios, entre eles a facilidade que um usuário sem conhecimentos profundos sobre computação tem de gerar e compartilhar resultados de uma pesquisa, pois os dados

são facilmente publicáveis. A interação dos usuários com os resultados é bastante dinâmica o que potencializa um abordagem *top-down*, em que o usuário a partir de dados genômicos consegue se aprofundar até encontrar informações genéticas que poderiam justificar suas hipóteses iniciais (com base na anotação fenotípica).

O GTACG foi proposto para auxiliar principalmente pesquisadores sem grandes conhecimentos sobre computação, permitindo a eles testarem hipóteses complexas sem que seja necessário qualquer tipo de programação adicional. Porém, para aqueles usuários mais especializados, o GTACG produz resultados em formatos abertos e facilmente reutilizáveis, assim como formatos compatíveis com o Roary. Para a fase final do ciclo de vida de uma pesquisa, os resultados visuais produzidos pelo arcabouço também visam a ajudar o pesquisador, fornecendo imagens em formatos com qualidade de publicação como o SVG, PNG ou TIFF.

5.1 Trabalho Futuros

Existe uma ampla gama de trabalhos que podem se beneficiar das bases desenvolvidas neste trabalho, entre eles destacam-se as análises sobre grupos de genomas. A enorme quantidade de dados gerados pela comparação de grupos de genomas não foi explorada em sua totalidade, podendo existir diversas extensões interessantes ainda a serem desenvolvidas. A maioria dos mecanismos de busca desenvolvidos no decorrer deste projeto se baseia em filtros lineares, abrindo margem principalmente para estudos baseados em filtros não lineares, como o uso de algumas técnicas de inteligência artificial.

A abordagem preferível para a busca de informações definidas no *website* gerado é a *top-down*, na qual o usuário parte de dados genômicos para se aprofundar em detalhes sobre dados genéticos, a partir das informações anotadas pelo usuário (fenotípicas ou genotípicas). Entretanto, caso a anotação dos genomas tenha sido feita de forma imprecisa, o restante do processo de busca pode ser prejudicado. Uma abordagem que interessante para diminuir os impactos desse problema é a utilização de uma abordagem *botton-up*, na qual a partir das famílias definidas poder-se-ia formular um método de sugestão de grupos genéticos, adaptando os grupos iniciais para serem mais condizentes com os resultados obtidos, ou sugerindo grupos de genomas inéditos (pouco intuitivos) que poderiam fomentar novos estudos.

Outro vasto campo de estudo ainda não explorado neste projeto são as diversas informações filogenéticas contidas nas famílias. Cada família de CDS representa, em última análise, uma filogenia, sendo várias delas conflitantes com a filogenia definida para os genomas. A análise de cada uma dessas filogenias poderia complementar o conhecimento sobre os eventos evolutivos que ocorreram sobre os genomas, sendo inclusive úteis para a identificação de famílias ou regiões que passaram por processos de recombinação ou de transferência horizontal.

Para facilitar o trabalho de usuários mais experientes, a interface de interação poderia ter uma alternativa em linha de comando, desde que fosse capaz de gerar todas as buscas que são feitas atualmente por meio do *website* estático. Essa interface poderia ser construída para executar apenas de forma *off-line*, porém ela seria mais útil caso o usuário tivesse a opção de se comunicar por linhas de comando em qualquer *website on-line* que faça o uso do GTACG.

Por fim, atualmente os estudos sobre as anotações fenotípicas podem ser bastante exaustivos, exigindo uma extensa investigação manual na literatura. Uma interessante adição a esse projeto seria tornar automático (ou ao menos semi-automático) esse processo por meio de mineração de texto com o objetivo de identificar nos documentos sobre os genomas (por exemplo, artigos científicos) quais apresentam ou não informações sobre a expressão do fenótipo de interesse.

5.2 Publicações relacionadas ao desenvolvimento da tese

Durante o desenvolvimento deste projeto foram publicados três artigos em revistas nacionais ou internacionais descrevendo resultados parciais desta pesquisa, listados a seguir:

1. SANTIAGO, C.; PEREIRA, V.; DIGIAMPIETRI, L. Homology Detection Using Multilayer Maximum Clustering Coefficient. *Journal of Computational Biology*, v. 25, n. 12, p. 1328–1338, 12 2018. ISSN 1557-8666. Disponível em: ⟨https://www.liebertpub.com/doi/10.1089/cmb.2017.0266⟩.
2. DIGIAMPIETRI, L. A. et al. A gene based bacterial whole genome comparison toolkit. *Revista de Informática Teórica e Aplicada*, v. 26, n. 1, p. 36, 4 2019. ISSN 21752745. Disponível em: ⟨https://seer.ufrgs.br/rita/article/view/RITA-VOL26-NR1-36⟩.

3. SANTIAGO, C. et al. Gene Tags Assessment by Comparative Genomics (GTACG): A user-friendly framework for bacterial comparative genomics. *Frontiers in Genetics*, 2019.

Referências[1]

ABASCAL, F.; VALENCIA, A. Clustering of proximal sequence space for the identification of protein families. *BIOINFORMATICS*, v. 18, n. 7, p. 908–921, 2002. Citado 2 vezes nas páginas 23 e 31.

APARNA, G. et al. A cell wall–degrading esterase of Xanthomonas oryzae requires a unique substrate recognition module for pathogenesis on rice. *The Plant Cell*, Am Soc Plant Biol, v. 21, n. 6, p. 1860–1873, 2009. Citado 2 vezes nas páginas 71 e 72.

ASSIS, R. d. A. B. et al. Identification and analysis of seven effector protein families with different adaptive and evolutionary histories in plant-associated members of the Xanthomonadaceae. *Scientific reports*, Nature Publishing Group, v. 7, n. 1, p. 16133, 2017. Citado 2 vezes nas páginas 71 e 72.

BELL, G.; HEY, T.; SZALAY, A. Beyond the data deluge. *Science*, American Association for the Advancement of Science, v. 323, n. 5919, p. 1297–1298, 2009. Citado na página 20.

BENEDICT, M. N. et al. ITEP: An integrated toolkit for exploration of microbial pan-genomes. *BMC Genomics*, v. 15, n. 1, p. 8, 2014. ISSN 1471-2164. Disponível em: ⟨http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-15-8⟩. Citado 2 vezes nas páginas 64 e 66.

BERGER, B.; PENG, J.; SINGH, M. Computational solutions for omics data. *Nature Reviews Genetics*, v. 14, n. 5, p. 333–346, 5 2013. ISSN 1471-0056. Disponível em: ⟨http://www.nature.com/articles/nrg3433⟩. Citado na página 21.

BLANVILLAIN, S. et al. Plant carbohydrate scavenging through TonB-dependent receptors: a feature shared by phytopathogenic and aquatic bacteria. *PLoS one*, Public Library of Science, v. 2, n. 2, p. e224, 2007. Citado 2 vezes nas páginas 72 e 73.

BORG, I.; GROENEN, P. J. F. *Modern multidimensional scaling: Theory and applications*. [S.l.]: Springer Science & Business Media, 2005. Citado na página 37.

BOULANGER, A. et al. The Plant Pathogen Xanthomonas campestris pv. campestris Exploits N-Acetylglucosamine during Infection. *mBio*, American Society for Microbiology, v. 5, n. 5, 2014. Disponível em: ⟨https://mbio.asm.org/content/5/5/e01527-14⟩. Citado na página 72.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Citado na página 65.

BREIMAN, R. F. et al. Defining the group A streptococcal toxic shock syndrome: rationale and consensus definition. *Jama*, American Medical Association, v. 269, n. 3, p. 390–391, 1993. Citado na página 56.

BRETTIN, T. et al. RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Scientific Reports*, v. 5, n. 1, p. 8365, 7 2015. ISSN 2045-2322. Disponível em: ⟨http://www.nature.com/articles/srep08365⟩. Citado na página 38.

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

BROHÉE, S.; HELDEN, J. van. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC bioinformatics*, v. 7, p. 488, 2006. ISSN 1471-2105. Citado 2 vezes nas páginas 23 e 39.

CAMACHO, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics*, v. 10, n. 1, p. 421, 2009. ISSN 1471-2105. Disponível em: ⟨http://www.biomedcentral.com/1471-2105/10/421⟩. Citado na página 31.

CARAPETIS, J. R. et al. The global burden of group A streptococcal diseases. *The Lancet infectious diseases*, Elsevier, v. 5, n. 11, p. 685–694, 2005. Citado na página 56.

CASADESUS, J.; LOW, D. Epigenetic Gene Regulation in the Bacterial World. *Microbiology and Molecular Biology Reviews*, v. 70, n. 3, p. 830–856, 9 2006. ISSN 1092-2172. Disponível em: ⟨http://mmbr.asm.org/cgi/doi/10.1128/MMBR.00016-06⟩. Citado na página 20.

CHAUDHARI, N. M.; GUPTA, V. K.; DUTTA, C. BPGA-an ultra-fast pan-genome analysis pipeline. *Scientific Reports*, Nature Publishing Group, v. 6, n. April, p. 1–10, 2016. ISSN 20452322. Disponível em: ⟨http://dx.doi.org/10.1038/srep24373⟩. Citado 3 vezes nas páginas 60, 63 e 64.

CHERVITZ, S. A. et al. Data Standards for Omics Data: The Basis of Data Sharing and Reuse. In: . [s.n.], 2011. p. 31–69. Disponível em: ⟨http://link.springer.com/10.1007/978-1-61779-027-0_2⟩. Citado 2 vezes nas páginas 20 e 21.

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, 10 2004. ISSN 0893-8512. Disponível em: ⟨http://cmr.asm.org/cgi/doi/10.1128/CMR.17.4.840-862.2004⟩. Citado na página 64.

COMIN, M.; VERZOTTO, D. Whole-genome phylogeny by virtue of unic subwords. *Proceedings - International Workshop on Database and Expert Systems Applications, DEXA*, p. 190–194, 2012. ISSN 15294188. Citado na página 22.

CONTRERAS-MOREIRA, B.; VINUESA, P. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Applied and Environmental Microbiology*, v. 79, n. 24, p. 7696–7701, 2013. ISSN 00992240. Citado 4 vezes nas páginas 23, 24, 39 e 66.

CORNEJO, O. E. et al. Evolutionary and Population Genomics of the Cavity Causing Bacteria Streptococcus mutans. *Molecular Biology and Evolution*, v. 30, n. 4, p. 881–893, 4 2013. ISSN 1537-1719. Disponível em: ⟨https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/mss278⟩. Citado na página 20.

CREEVEY, C. J.; MCINERNEY, J. O. Clann: Investigating phylogenetic information through supertree analyses. *Bioinformatics*, v. 21, n. 3, p. 390–392, 2005. ISSN 13674803. Citado 2 vezes nas páginas 36 e 104.

CREEVEY, C. J.; MCINERNEY, J. O. Trees from Trees: Construction of Phylogenetic Supertrees Using Clann. In: . [S.l.: s.n.], 2009. p. 139–161. Citado na página 25.

CUNNINGHAM, M. W. Pathogenesis of group A streptococcal infections. *Clinical microbiology reviews*, Am Soc Microbiol, v. 13, n. 3, p. 470–511, 2000. Citado na página 56.

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, 5 2005. ISSN 1471-0056. Disponível em: ⟨http://www.nature.com/articles/nrg1603⟩. Citado na página 24.

DIGIAMPIETRI, L. A. et al. A gene based bacterial whole genome comparison toolkit. *Revista de Informática Teórica e Aplicada*, v. 26, n. 1, p. 36, 4 2019. ISSN 21752745. Disponível em: ⟨https://seer.ufrgs.br/rita/article/view/RITA-VOL26-NR1-36⟩. Citado na página 77.

DING, W.; BAUMDICKER, F.; NEHER, R. A. panX: pan-genome analysis and exploration. *Nucleic Acids Research*, Oxford University Press, v. 46, n. 1, p. e5–e5, 1 2018. ISSN 0305-1048. Disponível em: ⟨http://academic.oup.com/nar/article/46/1/e5/4564799⟩. Citado 5 vezes nas páginas 23, 24, 35, 64 e 104.

DOBRINDT, U. et al. Genomic islands in pathogenic and environmental microorganisms. *Nature Reviews Microbiology*, v. 2, n. 5, p. 414–424, 5 2004. ISSN 1740-1526. Disponível em: ⟨http://www.nature.com/articles/nrmicro884⟩. Citado na página 69.

DONGEN, S. v. *Graph clustering by flow simulation*. Tese (Doutorado) — University of Utrecht, 2000. Citado na página 23.

DUPOIRON, S. et al. The N-Glycan Cluster from Xanthomonas campestris pv. campestris A TOOLBOX FOR SEQUENTIAL PLANT N-GLYCAN PROCESSING. *Journal of Biological Chemistry*, ASBMB, v. 290, n. 10, p. 6022–6036, 2015. Citado na página 72.

EDGAR, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics*, v. 5, p. 113, 8 2004. ISSN 1471-2105. Disponível em: ⟨http://www.ncbi.nlm.nih.gov/pubmed/15318951http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC517706⟩. Citado 2 vezes nas páginas 35 e 50.

ENELI, I.; DAVIES, H. D. Epidemiology and outcome of necrotizing fasciitis in children: an active surveillance study of the Canadian Paediatric Surveillance Program. *The Journal of pediatrics*, Elsevier, v. 151, n. 1, p. 79–84, 2007. Citado na página 56.

ENRIGHT, A. J.; DONGEN, S. V.; OUZOUNIS, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic acids research*, v. 30, n. 7, p. 1575–1584, 2002. ISSN 1362-4962. Citado 2 vezes nas páginas 23 e 31.

ENRIGHT, A. J.; OUZOUNIS, C. A. GeneRAGE: a robust algorithm for sequence clustering and domain detection. *BIOINFORMATICS*, v. 16, n. 5, p. 451–457, 2000. Citado na página 22.

Fa Zhang et al. Clustering orthologs based on sequence and domain similarities. In: *Eighth International Conference on High-Performance Computing in Asia-Pacific Region (HPCASIA'05)*. IEEE, 2005. p. 7 pp.–651. ISBN 0-7695-2486-9. Disponível em: ⟨http://ieeexplore.ieee.org/document/1592336/⟩. Citado na página 24.

FELSENSTEIN, J. Phylogenies from molecular sequences: inference and reliability. *Annual review of genetics*, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 22, n. 1, p. 521–565, 1988. Citado na página 20.

FELSENSTEIN, J. *PHYLIP (Phylogeny Inference Package) version 3.6*. [S.l.]: Department of Genome Sciences, University of Washington, Seattle., 2005. Citado 2 vezes nas páginas 35 e 105.

FIETTO, J. L. R.; MACIEL, T. E. F. Sequenciando genomas. In: MOREIRA, L. M. (Ed.). *Ciências genômicas: fundamentos e aplicações*. [S.l.]: Sociedade Brasileira de Computação, 2015. p. 27–64. ISBN 978-85-89265-22-5. Citado na página 20.

FIETTO, L. G.; LAMËGO, M. R. d. A. História e importância da genômica. In: MOREIRA, L. M. (Ed.). *Ciências genômicas: fundamentos e aplicações*. [S.l.]: Sociedade Brasileira de Computação, 2015. p. 21–26. ISBN 978-85-89265-22-5. Citado na página 20.

GROUP, A. *MDSJ: Java Library for Multidimensional Scaling*. University of Konstanz, 2009. Disponível em: ⟨http://www.inf.uni-konstanz.de/algo/software/mdsj/⟩. Citado na página 105.

GUINDON, S. et al. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, v. 59, n. 3, p. 307–321, 3 2010. ISSN 1076-836X. Disponível em: ⟨https://academic.oup.com/sysbio/article/59/3/307/1702850⟩. Citado 2 vezes nas páginas 35 e 50.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques*. Third edit. [S.l.]: Elsevier, 2011. 740 p. ISBN 978-0-12-381479-1. Citado na página 22.

HARDISON, R. C. Comparative Genomics. *PLoS Biol*, Public Library of Science, v. 1, n. 2, 2003. Disponível em: ⟨http://dx.doi.org/10.1371/journal.pbio.0000058⟩. Citado na página 20.

HAUBEN, L. et al. Comparison of 16S Ribosomal DNA Sequences of All Xanthomonas Species. *International Journal of Systematic Bacteriology*, v. 47, n. 2, p. 328–335, 4 1997. ISSN 0020-7713. Disponível em: ⟨http://ijs.microbiologyresearch.org/content/journal/ijsem/10.1099/00207713-47-2-328⟩. Citado na página 25.

HÄWEKER, H. et al. Pattern Recognition Receptors Require N-Glycosylation to Mediate Plant Immunity. *Journal of Biological Chemistry*, v. 285, n. 7, p. 4629–4636, 2010. Disponível em: ⟨http://www.jbc.org/content/285/7/4629.abstract⟩. Citado na página 71.

HILLMER, R. A. Systems biology for biologists. *PLoS pathogens*, Public Library of Science, v. 11, n. 5, p. e1004786, 2015. Citado na página 66.

HOLLINGSHEAD, S. K. et al. Molecular evolution of a multigene family in group A streptococci. *Molecular biology and evolution*, v. 11, n. 2, p. 208–219, 1994. Citado na página 69.

ILINA, E. N. et al. Comparative Genomic Analysis of Mycobacterium tuberculosis Drug Resistant Strains from Russia. *PLoS ONE*, v. 8, n. 2, p. e56577, 2 2013. ISSN 1932-6203. Disponível em: ⟨http://dx.plos.org/10.1371/journal.pone.0056577⟩. Citado 2 vezes nas páginas 20 e 64.

JALAN, N. U. *Comparative Genomic and Transcriptomic Analyses of Xanthomonas Citri Subsp. Citri and Related Species Provides Insights into Virulence and Host-Specificity*. Tese (Doutorado) — University of Florida, 2012. Citado na página 72.

JOYCE, E. A. et al. Redefining bacterial populations: a post-genomic reformation. *Nature Reviews Genetics*, v. 3, n. 6, p. 462–473, 6 2002. ISSN 1471-0056. Disponível em: ⟨http://www.nature.com/articles/nrg820⟩. Citado na página 20.

KEHDY, F. S. G. et al. Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. *Proceedings of the National Academy of Sciences*, v. 112, n. 28, p. 8696–8701, 7 2015. ISSN 0027-8424. Disponível em: ⟨http://www.pnas.org/lookup/doi/10.1073/pnas.1504447112⟩. Citado na página 21.

KOBOUROV, S. G. Spring Embedders and Force-Directed Graph Drawing Algorithms. 2012. *URL: http://arxiv.org/abs/1201.3011*, 2012. Citado na página 37.

LAIA, M. L. et al. New genes of Xanthomonas citri subsp. citri involved in pathogenesis and adaptation revealed by a transposon-based mutant library. *BMC microbiology*, BioMed Central, v. 9, n. 1, p. 12, 2009. Citado 2 vezes nas páginas 72 e 73.

LAING, C. et al. Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. *BMC bioinformatics*, BioMed Central, v. 11, n. 1, p. 461, 2010. Citado 2 vezes nas páginas 64 e 66.

LAMAGNI, T. L. et al. Epidemiology of Severe Streptococcus pyogenes Disease in Europe. *Journal of Clinical Microbiology*, v. 46, n. 7, p. 2359–2367, 7 2008. ISSN 0095-1137. Disponível em: ⟨http://jcm.asm.org/cgi/doi/10.1128/JCM.00422-08⟩. Citado na página 56.

LANCEFIELD, R. C.; PERLMANN, G. E. Preparation and properties of type-specific M antigen isolated from a group A, type 1 hemolytic streptococcus. *Journal of Experimental Medicine*, Rockefeller University Press, v. 96, n. 1, p. 71–82, 1952. Citado na página 56.

LANDER, E. S. et al. Initial sequencing and analysis of the human genome. *Nature*, v. 409, n. 6822, p. 860–921, 2 2001. ISSN 0028-0836. Disponível em: ⟨http://www.nature.com/doifinder/10.1038/35057062⟩. Citado na página 21.

LEE, N.-Y. et al. Clinical and Economic Impact of Multidrug Resistance in Nosocomial Acinetobacter baumannii Bacteremia. *Infection Control & Hospital Epidemiology*, v. 28, n. 6, p. 713–719, 6 2007. ISSN 0899-823X. Disponível em: ⟨https://www.cambridge.org/core/product/identifier/S0195941700046531/type/journal_article⟩. Citado na página 20.

LEIMEISTER, C.-A. et al. Fast alignment-free sequence comparison using spaced-word frequencies. *Bioinformatics*, v. 30, n. 14, p. 1991–1999, 7 2014. ISSN 1460-2059. Disponível em: ⟨https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu177⟩. Citado na página 25.

LI, L. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Research*, v. 13, n. 9, p. 2178–2189, 9 2003. ISSN 1088-9051. Disponível em: ⟨http://dx.doi.org/10.1101/gr.1224503http://www.genome.org/cgi/doi/10.1101/gr.1224503⟩. Citado 3 vezes nas páginas 23, 24 e 39.

LIN, C.-H. et al. Characterization of Xanthomonas campestris pv. campestris heat shock protein A (HspA), which possesses an intrinsic ability to reactivate inactivated proteins. *Applied microbiology and biotechnology*, Springer, v. 88, n. 3, p. 699–709, 2010. Citado 2 vezes nas páginas 71 e 72.

LUNAK, Z. R.; NOEL, K. D. A quinol oxidase, encoded by cyoABCD, is utilized to adapt to lower O2 concentrations in Rhizobium etli CFN42. *Microbiology*, Microbiology Society, v. 161, n. Pt 1, p. 203, 2015. Citado 2 vezes nas páginas 72 e 73.

MANSFIELD, J. et al. Top 10 plant pathogenic bacteria in molecular plant pathology. *Molecular Plant Pathology*, v. 13, n. 6, p. 614–629, 8 2012. ISSN 14646722. Disponível em: ⟨http://doi.wiley.com/10.1111/j.1364-3703.2012.00804.x⟩. Citado na página 20.

MENENDEZ, A.; FINLAY, B. B. Defensins in the immunology of bacterial infections. *Current Opinion in Immunology*, v. 19, n. 4, p. 385–391, 8 2007. ISSN 09527915. Disponível em: ⟨https://linkinghub.elsevier.com/retrieve/pii/S0952791507001124⟩. Citado na página 69.

MOREIRA, L. M. et al. Proteomics-based identification of differentially abundant proteins reveals adaptation mechanisms of Xanthomonas citri subsp. citri during Citrus sinensis infection. *BMC microbiology*, BioMed Central, v. 17, n. 1, p. 155, 2017. Citado 2 vezes nas páginas 72 e 73.

NASCIMENTO, R. et al. The type II secreted lipase/esterase LesA is a key virulence factor required for Xylella fastidiosa pathogenesis in grapevines. *Scientific reports*, Nature Publishing Group, v. 6, p. 18598, 2016. Citado 2 vezes nas páginas 71 e 72.

NAUSHAD, H. S.; GUPTA, R. S. Phylogenomics and molecular signatures for species from the plant pathogen-containing order Xanthomonadales. *PLoS One*, Public Library of Science, v. 8, n. 2, p. e55216, 2013. Citado na página 53.

NOVERR, M. C.; HUFFNAGLE, G. B. Does the microbiota regulate immune responses outside the gut? *Trends in Microbiology*, v. 12, n. 12, p. 562–568, 12 2004. ISSN 0966842X. Disponível em: ⟨https://linkinghub.elsevier.com/retrieve/pii/S0966842X04002409⟩. Citado na página 69.

OBOLSKI, U. et al. Identifying Streptococcus pneumoniae genes associated with invasive disease using pangenome-based whole genome sequence typing. 2018. Citado 3 vezes nas páginas 20, 64 e 65.

O'BRIEN, K. P.; REMM, M.; SONNHAMMER, E. L. L. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Research*, 2005. Citado na página 24.

PAGE, A. J. et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, v. 31, n. 22, p. 3691–3693, 2015. Disponível em: ⟨http://dx.doi.org/10.1093/bioinformatics/btv421⟩. Citado 7 vezes nas páginas 23, 24, 39, 47, 60, 63 e 64.

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. ISSN 1471-2164. Citado na página 67.

PRASANNA, A. N.; MEHRA, S. Comparative Phylogenomics of Pathogenic and Non-Pathogenic Mycobacterium. *PLoS ONE*, v. 8, n. 8, 2013. ISSN 19326203. Citado na página 64.

PRICE, M. N.; DEHAL, P. S.; ARKIN, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE*, v. 5, n. 3, p. e9490, 3 2010. ISSN 1932-6203. Disponível em: ⟨https://dx.plos.org/10.1371/journal.pone.0009490⟩. Citado 3 vezes nas páginas 24, 35 e 50.

RAMOS, P. L. et al. An MLSA-based online scheme for the rapid identification of Stenotrophomonas isolates. *Memórias do Instituto Oswaldo Cruz*, SciELO Brasil, v. 106, n. 4, p. 394–399, 2011. Citado na página 53.

SAKATA, H. Susceptibility and emm type of Streptococcus pyogenes isolated from children with severe infection. *Journal of Infection and Chemotherapy*, Springer, v. 19, n. 6, p. 1042–1046, 2013. Citado na página 56.

SANGER, F.; NICKLEN, S.; COULSON, A. R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, v. 74, n. 12, p. 5463–5467, 12 1977. ISSN 0027-8424 (Print). Citado na página 20.

SANTIAGO, C. et al. Gene Tags Assessment by Comparative Genomics (GTACG): A user-friendly framework for bacterial comparative genomics. *Frontiers in Genetics*, 2019. Citado 13 vezes nas páginas 29, 49, 52, 54, 61, 62, 63, 65, 68, 72, 73, 78 e 99.

SANTIAGO, C.; PEREIRA, V.; DIGIAMPIETRI, L. Homology Detection Using Multilayer Maximum Clustering Coefficient. *Journal of Computational Biology*, v. 25, n. 12, p. 1328–1338, 12 2018. ISSN 1557-8666. Disponível em: ⟨https://www.liebertpub.com/doi/10.1089/cmb.2017.0266⟩. Citado 8 vezes nas páginas 33, 41, 42, 43, 44, 45, 46 e 77.

SASSON, O.; LINIAL, N.; LINIAL, M. The metric space of proteins– comparative study of clustering algorithms. *BIOINFORMATICS*, v. 18, n. 1, p. 14–21, 2002. Disponível em: ⟨http://www.protonet.cs.huji.ac.il/examples.html.⟩ Citado 3 vezes nas páginas 23, 28 e 31.

SETUBAL, J. C.; STOYE, J.; STADLER, P. F. (Ed.). *Comparative genomics*. [S.l.: s.n.], 2018. v. 1704. 363–400 p. ISSN 10643745. ISBN 978-1-4939-7463-4. Citado na página 24.

SETUBAL, J. C.; WATTAM, R.; ALMEIDA, N. Comparative Genomics for Prokaryotes. In: *Methods in molecular biology*. [S.l.: s.n.], 2018. cap. 3. Citado na página 22.

SHARMA, V.; PATIL, P. B. Resolving the phylogenetic and taxonomic relationship of Xanthomonas and Stenotrophomonas strains using complete rpoB gene sequence. *PLoS currents*, Public Library of Science, v. 3, 2011. Citado na página 53.

SIEVERS, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology*, v. 7, n. 1, p. 539, 2011. ISSN 1744-4292. Disponível em: ⟨http://msb.embopress.org/content/7/1/539.abstract⟩. Citado 2 vezes nas páginas 35 e 50.

SIMMONS, S. L. et al. Population Genomic Analysis of Strain Variation in Leptospirillum Group II Bacteria Involved in Acid Mine Drainage Formation. *PLoS Biology*, v. 6, n. 7, p. e177, 7 2008. ISSN 1545-7885. Disponível em: ⟨https://dx.plos.org/10.1371/journal.pbio.0060177⟩. Citado na página 20.

SIMõES, S. N. et al. NERI: network-medicine based integrative approach for disease gene prioritization by relative importance. *BMC Bioinformatics*, v. 16, n. Suppl 19, p. S9, 2015. ISSN 1471-2105. Disponível em: ⟨http://dx.doi.org/10.1186/1471-2105-16-S19-S9http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-16-S19-S9⟩. Citado na página 20.

STAMATAKIS, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, Oxford University Press, v. 30, n. 9, p. 1312–1313, 2014. Citado na página 50.

STEINEGGER, M.; SÖDING, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, v. 35, n. 11, p. 1026–1028, 11 2017. ISSN 1087-0156. Disponível em: ⟨http://www.nature.com/articles/nbt.3988⟩. Citado 2 vezes nas páginas 31 e 62.

TAMAYO, E. et al. Streptococcus pyogenes pneumonia in adults: clinical presentation and molecular characterization of isolates 2006-2015. *PLoS One*, Public Library of Science, v. 11, n. 3, p. e0152640, 2016. Citado na página 56.

TETTELIN, H. et al. Comparative genomics: the bacterial pan-genome. *Current Opinion in Microbiology*, v. 11, n. 5, p. 472–477, 10 2008. ISSN 13695274. Disponível em: ⟨https://linkinghub.elsevier.com/retrieve/pii/S1369527408001239⟩. Citado na página 24.

VERNIKOS, G. et al. Ten years of pan-genome analyses. *Current Opinion in Microbiology*, v. 23, p. 148–154, 2015. ISSN 18790364. Citado na página 64.

VLIET, A. H. M. van. Use of pan-genome analysis for the identification of lineage-specific genes of Helicobacter pylori. *FEMS microbiology letters*, Oxford University Press, v. 364, n. 2, 2017. Citado na página 64.

VOGEL, C. et al. Structure, function and evolution of multidomain proteins. *Current Opinion in Structural Biology*, v. 14, n. 2, p. 208–216, 2004. ISSN 0959440X. Citado 2 vezes nas páginas 33 e 45.

WATTAM, A. R. et al. Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. *Nucleic Acids Research*, v. 45, n. D1, p. D535–D542, 1 2017. ISSN 0305-1048. Disponível em: ⟨https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw1017⟩. Citado 2 vezes nas páginas 38 e 103.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world'networks. *nature*, Nature Publishing Group, v. 393, n. 6684, p. 440, 1998. Citado na página 30.

XIA, X. *Comparative Genomics*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. (SpringerBriefs in Genetics). ISBN 978-3-642-37145-5. Disponível em: ⟨http://link.springer.com/10.1007/978-3-642-37146-2⟩. Citado na página 20.

YACHDAV, G. et al. MSAViewer: interactive JavaScript visualization of multiple sequence alignments. *Bioinformatics*, Oxford University Press, v. 32, n. 22, p. 3501–3503, 2016. Citado na página 50.

ZHAO, Y. et al. PanGP: a tool for quickly analyzing bacterial pan-genome profile. *Bioinformatics*, Oxford University Press, v. 30, n. 9, p. 1297–1299, 2014. Citado 3 vezes nas páginas 60, 63 e 64.

ZHAO, Y. et al. PGAP-X: extension on pan-genome analysis pipeline. *BMC genomics*, BioMed Central, v. 19, n. 1, p. 36, 2018. Citado 2 vezes nas páginas 64 e 66.

ZHAO, Y. et al. PGAP: pan-genomes analysis pipeline. *Bioinformatics*, v. 28, n. 3, p. 416–418, 2 2012. ISSN 1460-2059. Disponível em: ⟨https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr655⟩. Citado na página 66.

ZHOU, X.; YAN, Q.; WANG, N. Deciphering the regulon of a GntR family regulator via transcriptome and ChIP-exo analyses and its contribution to virulence in Xanthomonas citri. *Molecular Plant Pathology*, v. 18, n. 2, p. 249–262, 2 2017. ISSN 14646722. Disponível em: ⟨http://doi.wiley.com/10.1111/mpp.12397⟩. Citado na página 72.

ZHOU, X.; YAN, Q.; WANG, N. Deciphering the regulon of a GntR family regulator via transcriptome and ChIP-exo analyses and its contribution to virulence in Xanthomonas citri. *Molecular Plant Pathology*, v. 18, n. 2, p. 249–262, 2017. Disponível em: ⟨https://onlinelibrary.wiley.com/doi/abs/10.1111/mpp.12397⟩. Citado na página 71.

Anexo A – Dados genômicos utilizados nos estudos de caso

A.1 Genomas de Streptococcus pyogenes

Quadro 3 – Informações sobre os 55 genomas de *Streptococcus pyogenes* que foram utilizados nos estudos de caso, incluindo o código de acesso para o genoma no NCBI.

Acesso	Nome	Genótipo *emm*	Invasividade	Padrão	País
gi|703558587	1E1	44	U	E	FR
gi|828455247	5448	1	I	A-C	GER
gi|409692283	A20	1	I	A-C	TAI
gi|818416626	AP1	1	I	A-C	CZE
gi|1020263163	AP53	53	N	D	USA
gi|749295042	ATCC 19615	80	N	D	USA
gi|386361880	Alab49	53	N	D	USA
gi|825741578	D471	6	N	A-C	USA
gi|920656811	H293	89	I	E	UK
gi|387932825	HKU16	12	N	A-C	CHN
gi|703570643	HKU360	12	N	A-C	AUS
gi|874011340	HKU488	1	N	A-C	AUS
gi|523444678	HSC5	14	I	A-C	USA
gi|1060084065	JMUB1235	89	I	E	JP
gi|823683938	JRS4	6	N	A-C	USA
gi|520190261	M1-476	1	I	A-C	JP
gi|686514231	M23ND	23	I	A-C	USA
gi|1041941603	M28PF1	28	I	E	FR
gi|1001622312	M3-b	3	I	A-C	JP
gi|982534187	MEW123	28	N	E	USA
gi|982532632	MEW427	4	N	E	USA
gi|94989509	MGAS10270	2	N	E	USA
gi|50902420	MGAS10394	6	N	A-C	USA
gi|94993396	MGAS10750	4	N	E	USA
gi|1024855856	MGAS11027	89	N	E	USA
gi|383479207	MGAS15252	59	I	D	USA
gi|378928860	MGAS1882	59	N	D	USA
gi|94991497	MGAS2096	12	N	A-C	USA
gi|1024795854	MGAS23530	89	N	E	USA
gi|1024852152	MGAS27061	89	I	E	USA
gi|21909536	MGAS315	3	I	A-C	USA
gi|861564765	MGAS5005	1	I	A-C	USA
gi|71902667	MGAS6180	28	I	E	USA
gi|19745201	MGAS8232	18	N	A-C	USA
gi|94541139	MGAS9429	12	N	A-C	USA
gi|760873924	MTB313	1	I	A-C	JP

(Continuação)

Acesso	Nome	Genótipo *emm*	Invasividade	Padrão	País
gi|760875820	MTB314	1	I	A-C	JP
gi|139472888	Manfredo	5	N	A-C	UK
gi|777206994	NCTC8198	1	N	A-C	UK
gi|917641723	NGAS322	114	I	E	CAN
gi|827378376	NGAS327	83	I	D	CAN
gi|827376749	NGAS596	82	I	E	CAN
gi|917643905	NGAS638	101	I	D	CAN
gi|827374941	NGAS743	87	I	E	CAN
gi|1026248336	NS53	71	N	D	USA
gi|209539788	NZ131	49	N	E	USA
gi|602625715	SF370	1	U	A-C	USA
gi|47118313	SSI-1	3	I	A-C	JP
gi|1041930325	STAB09014	28	N	E	FR
gi|836556487	STAB10015	28	N	E	FR
gi|749295047	STAB1101	83	I	D	FR
gi|666903168	STAB1102	83	I	D	FR
gi|1047888374	STAB13021	66	I	E	FR
gi|666904753	STAB901	44	I	E	FR
gi|755007402	STAB902	3	I	A-C	FR

Fonte: Caio Santiago, 2019

A.2 Informações relacionadas às doenças causadas pelos Streptococcus pyogenes

Quadro 4 – Informações sobre as doenças causadas pelos 55 genomas de *Streptococcus pyogenes* utilizados nos estudos de casos.

Acesso	Nome	Doença
gi|703558587	1E1	U
gi|828455247	5448	Necrotizing Fasciitis
gi|409692283	A20	Necrotizing Fasciitis
gi|818416626	AP1	U
gi|1020263163	AP53	Impetigo
gi|749295042	ATCC 19615	Pharyngitis
gi|386361880	Alab49	Impetigo
gi|825741578	D471	Acute Rheumatic Fever
gi|920656811	H293	Necrotizing Fasciitis
gi|387932825	HKU16	Scarlet Fever
gi|703570643	HKU360	Scarlet Fever
gi|874011340	HKU488	Scarlet Fever
gi|523444678	HSC5	U
gi|1060084065	JMUB1235	Acute Phlegmonous Gastritis

(Continuação)

Acesso	Nome	Doença
gi|823683938	JRS4	Acute Rheumatic Fever
gi|520190261	M1-476	Streptococcal Toxic Shock Syndrome
gi|686514231	M23ND	Necrotizing Fasciitis
gi|1041941603	M28PF1	Endometritis
gi|1001622312	M3-b	Streptococcal Toxic Shock Syndrome
gi|982534187	MEW123	Pharyngitis
gi|982532632	MEW427	Pharyngitis
gi|94989509	MGAS10270	Superficial Dermatitis
gi|50902420	MGAS10394	Pharyngitis
gi|94993396	MGAS10750	Pharyngitis
gi|1024855856	MGAS11027	Pharyngitis
gi|383479207	MGAS15252	Soft Tissue Infection
gi|378928860	MGAS1882	Acute Poststreptococcal Glomerulonephritis
gi|94991497	MGAS2096	Acute Poststreptococcal Glomerulonephritis
gi|1024795854	MGAS23530	Pharyngitis
gi|1024852152	MGAS27061	U
gi|21909536	MGAS315	Pharyngitis
gi|861564765	MGAS5005	Cerebrospinal Fluid Infection
gi|71902667	MGAS6180	Puerperal Sepsis
gi|19745201	MGAS8232	Acute Rheumatic Fever
gi|94541139	MGAS9429	Pharyngitis
gi|760873924	MTB313	Meningitis
gi|760875820	MTB314	Meningitis
gi|139472888	Manfredo	Acute Rheumatic Fever
gi|777206994	NCTC8198	Scarlet Fever
gi|917641723	NGAS322	Bacteremia
gi|827378376	NGAS327	Bacteremia
gi|827376749	NGAS596	Bacteremia
gi|917643905	NGAS638	Bacteremia
gi|827374941	NGAS743	Necrotizing Fasciitis
gi|1026248336	NS53	Skin Infection
gi|209539788	NZ131	Acute Poststreptococcal Glomerulonephritis
gi|602625715	SF370	Wound Infection
gi|47118313	SSI-1	Streptococcal Toxic Shock Syndrome
gi|1041930325	STAB09014	Perianal Streptococcal Cellulitis

(Continuação)

Acesso	Nome	Doença
gi|836556487	STAB10015	PERIANAL STREPTOCOCCAL CELLULITIS
gi|749295047	STAB1101	NECROTIZING FASCIITIS
gi|666903168	STAB1102	NECROTIZING FASCIITIS
gi|1047888374	STAB13021	SUBCUTANEOUS ABSCESS
gi|666904753	STAB901	ENDOMETRITIS
gi|755007402	STAB902	STREPTOCOCCAL TOXIC SHOCK SYNDROME

Fonte: Caio Santiago, 2019

A.3 Genomas de Xanthomonadaceae

Quadro 5 – Informações sobre os 161 genomas da família *Xanthomonadaceae* que foram utilizados nos estudos de caso, incluindo o código de acesso para o genoma no NCBI.

Acesso	Nome	Abreviação	Fito-patógeno	Associado a plantas	Usado no teste com 10 genomas	Usado no teste com 20 genomas	Usado no teste com 30 genomas	Usado no teste com 40 genomas	Usado no teste com 50 genomas	Usado no teste com 69 genomas
GCA_000233915.4	*PseudoXanthomonas spadix* BD-a59	PspadixBD-a59								X
GCA_000185965.1	*PseudoXanthomonas suwonensis* 11-1	Psuwon11-1								X
GCA_000972865.1	*Pseudoxanthomonas suwonensis* strain J1	PsuwonJ1								X
GCA_001314305.1	*Stenotrophomonas acidaminiphila* strain ZAC14D2 NAIMI4 2	Sacidaminiphila								X
GCA_002025605.1	*Stenotrophomonas maltophilia* AA1	SmaltAA1								
GCA_002847385.1	*Stenotrophomonas maltophilia* CSM2	SmaltCSM2								
GCA_000284595.1	*Stenotrophomonas maltophilia* D457	SmaltD457								X
GCA_001071475.1	*Stenotrophomonas maltophilia* FDAARGOS 325	SmaltFDAARGOS325								
GCA_002951115.1	*Stenotrophomonas maltophilia* FDAARGOS 92	SmaltFDAARGOS92								

(Continuação)

Acesso	Nome	Abreviação	Fito-patógeno	Associado a plantas	Usado no teste com 10 genomas	Usado no teste com 20 genomas	Usado no teste com 30 genomas	Usado no teste com 40 genomas	Usado no teste com 50 genomas	Usado no teste com 69 genomas
GCA_000223885.1	*Stenotrophomonas maltophilia* JV3	SmaltJV3								X
GCA_000072485.1	*Stenotrophomonas maltophilia* K279a strain K279a	SmaltK279a								X
GCA_900186865.1	*Stenotrophomonas maltophilia* NCTC10257	SmaltNCTC10257								
GCA_002138415.1	*Stenotrophomonas maltophilia* OUC Est10	SmaltOUCEst10								
GCA_000020665.1	*Stenotrophomonas maltophilia* R551-3	SmaltR551-3								X
GCA_001274655.1	*Stenotrophomonas maltophilia* strain ISMMS2	SmaltISMMS2								X
GCA_001274675.1	*Stenotrophomonas maltophilia* strain ISMMS2R	SmaltISMMS2R								X
GCA_001274595.1	*Stenotrophomonas maltophilia* strain ISMMS3	SmaltISMMS3								X
GCA_001700965.1	*Stenotrophomonas nitritireducens* 2001	Snitrit2001								
GCA_001704155.1	*Stenotrophomonas rhizophila* QL-P4	SrhizophilaQL-P4								
GCA_002192255.1	*Stenotrophomonas* sp. *WZN-1*	SWZN-1WZN-1								
GCA_001562215.1	*Stenotrophomonas* sp. KCTC 12332 YM1	SKCTC12332YM1								
GCA_001806305.1	*Stenotrophomonas* sp. LM091	SLM091LM091								
GCA_000087965.1	*Xanthomonas albilineans* GPE PC73	XalbGPEPC73	X	X	X	X	X	X	X	X
GCA_000007165.1	*Xanthomonas axonopodis* pv. *citri* str. 306	Xaxcitri306	X	X	X	X	X	X	X	X
GCA_000225915.1	*Xanthomonas axonopodis* pv. *citrumelo* F1	XaxcitrumeloF1	X	X	X	X	X	X	X	X
GCA_000348585.1	*Xanthomonas axonopodis* Xac29-1	XaxXac29-1	X	X	X	X	X	X	X	X
GCA_002879955.1	*Xanthomonas campestris* pv. *campestris* 3811	Xcc3811	X	X	X	X	X	X	X	
GCA_000012105.1	*Xanthomonas campestris* pv. *campestris* str. 8004	Xcc8004	X	X	X	X	X	X	X	X
GCA_000007145.1	*Xanthomonas campestris* pv. *campestris* str. ATCC 33913	XccATCC33913	X	X	X	X	X	X	X	X

(Continuação)

Acesso	Nome	Abreviação	Fito-patógeno	Associado a plantas	Usado no teste com 10 genomas	Usado no teste com 20 genomas	Usado no teste com 30 genomas	Usado no teste com 40 genomas	Usado no teste com 50 genomas	Usado no teste com 69 genomas
GCA_002776775.1	*Xanthomonas campestris* pv. *campestris* str. CN12	XccCN12CN12	X	X	X	X	X	X	X	
GCA_002776835.1	*Xanthomonas campestris* pv. *campestris* str. CN18	XccCN18CN18	X	X	X	X	X	X	X	
GCA_000070605.1	*Xanthomonas campestris* pv. *campestris* strain B100	XccB100	X	X	X	X	X	X	X	X
GCA_001186415.1	*Xanthomonas campestris* pv. *campestris* strain ICMP 21080	XccICMP21080	X	X		X	X	X	X	X
GCA_001186465.1	*Xanthomonas campestris* pv. *campestris* strain ICMP 4013	XccICMP4013	X	X		X	X	X	X	X
GCA_000221965.1	*Xanthomonas campestris* pv. *raphani* 756C	Xcraphani756C	X	X		X	X	X	X	X
GCA_000009165.1	*Xanthomonas campestris* pv. *vesicatoria*	Xcvesicatoria	X	X		X	X	X	X	X
GCA_000009165.1	*Xanthomonas campestris* pv. *vesicatoria* str. 85-10	Xcvesicatoria85-10	X	X		X	X	X	X	X
GCA_000972745.1	*Xanthomonas campestris* strain 17	Xcampestris17	X	X		X	X	X	X	X
GCA_001028285.3	*Xanthomonas citri* pv. *citri* strain jx-6	Xccitrijx-6	X	X		X	X	X	X	X
GCA_002163775.1	*Xanthomonas citri* pv. *glycines* str. 12-2	Xcglycines12-2	X	X		X	X	X	X	
GCA_001854145.2	*Xanthomonas citri* pv. *glycines* str. 8ra	Xcglycines8ra	X	X		X	X	X	X	
GCA_002240395.1	*Xanthomonas citri* pv. *mangiferaeindicae*	Xcmangifer	X	X		X	X	X	X	
GCA_002759275.1	*Xanthomonas citri* pv. *phaseoli* var. fuscans CFBP6988R	XcpfuscansCFBP6988R	X	X			X	X	X	
GCA_002759355.1	*Xanthomonas citri* pv. *phaseoli* var. *fuscans* CFBP4885	XcpfuscansCFBP4885	X	X			X	X	X	
GCA_002759215.1	*Xanthomonas citri* pv. *phaseoli* var. *fuscans* CFBP6165	XcpfuscansCFBP6165	X	X			X	X	X	
GCA_002759235.1	*Xanthomonas citri* pv. *phaseoli* var. *fuscans* CFBP6166	XcpfuscansCFBP6166	X	X			X	X	X	

(Continuação)

Acesso	Nome	Abreviação	Fito-patógeno	Associado a plantas	Usado no teste com 10 genomas	Usado no teste com 20 genomas	Usado no teste com 30 genomas	Usado no teste com 40 genomas	Usado no teste com 50 genomas	Usado no teste com 69 genomas
GCA_002759415.1	*Xanthomonas citri* pv. *phaseoli* var. *fuscans* CFBP6167	XcpfuscansCFBP6167	X	X			X	X	X	
GCA_002759255.1	*Xanthomonas citri* pv. *phaseoli* var. *fuscans* CFBP6975	XcpfuscansCFBP6975	X	X			X	X	X	
	Xanthomonas citri pv. *phaseoli* var. *fuscans* CFBP6989	XcpfuscansCFBP6989	X	X			X	X	X	
GCA_002759315.1	*Xanthomonas citri* pv. *phaseoli* var. *fuscans* CFBP6990	XcpfuscansCFBP6990	X	X			X	X	X	
GCA_002759395.1	*Xanthomonas citri* pv. *phaseoli* var. *fuscans* CFBP6991	XcpfuscansCFBP6991	X	X			X	X	X	
GCA_002759335.1	*Xanthomonas citri* pv. *phaseoli* var. *fuscans* CFBP6992	XcpfuscansCFBP6992	X	X			X	X	X	
GCA_002759175.1	*Xanthomonas citri* pv. *phaseoli* var. *fuscans* CFBP6994R	XcpfuscansCFBP6994R	X	X				X	X	
GCA_002759195.1	*Xanthomonas citri* pv. *phaseoli* var. *fuscans* CFBP6996R	XcpfuscansCFBP6996R	X	X				X	X	
GCA_002759375.1	*Xanthomonas citri* pv. *phaseoli* var. *fuscans* CFBP7767	XcpfuscansCFBP7767	X	X				X	X	
GCA_002218245.1	*Xanthomonas citri* pv. *vignicola* CFBP7111	XcvignicolaCFBP7111	X	X				X	X	
GCA_002218265.1	*Xanthomonas citri* pv. *vignicola* CFBP7112	XcvignicolaCFBP7112	X	X				X	X	
GCA_002218285.1	*Xanthomonas citri* pv. *vignicola* CFBP7113	XcvignicolaCFBP7113	X	X				X	X	
GCA_000816885.1	*Xanthomonas citri* subsp. *citri* A306	XccA306	X	X				X	X	X
GCA_000349225.1	*Xanthomonas citri* subsp. *citri* Aw12879	XccAw12879	X	X				X	X	X
GCA_001922105.1	*Xanthomonas citri* subsp. *citri* LH201	XccLH201	X	X				X	X	
GCA_001922065.1	*Xanthomonas citri* subsp. *citri* LH276	XccLH276	X	X				X	X	
GCA_001922085.1	*Xanthomonas citri* subsp. *citri* LJ207-7	XccLJ207-7	X	X					X	
GCA_001922045.1	*Xanthomonas citri* subsp. *citri* LL074-4	XccLL074-4	X	X					X	

(Continuação)

Acesso	Nome	Abreviação	Fito-patógeno	Associado a plantas	Usado no teste com 10 genomas	Usado no teste com 20 genomas	Usado no teste com 30 genomas	Usado no teste com 40 genomas	Usado no teste com 50 genomas	Usado no teste com 69 genomas
GCA_000961415.1	*Xanthomonas citri* subsp. *citri* strain 5208	Xcc5208	X	X					X	X
GCA_000961435.1	*Xanthomonas citri* subsp. *citri* strain AW13	XccAW13	X	X					X	X
GCA_000961455.1	*Xanthomonas citri* subsp. *citri* strain AW14	XccAW14	X	X					X	X
GCA_000961475.1	*Xanthomonas citri* subsp. *citri* strain AW15	XccAW15	X	X					X	X
GCA_000961495.1	*Xanthomonas citri* subsp. *citri* strain AW16	XccAW16	X	X					X	X
GCA_000961395.1	*Xanthomonas citri* subsp. *citri* strain BL18	XccBL18	X	X					X	X
GCA_000961375.1	*Xanthomonas citri* subsp. *citri* strain FB19	XccFB19	X	X					X	X
GCA_000961355.1	*Xanthomonas citri* subsp. *citri* strain gd2	Xccgd2	X	X					X	X
GCA_002759095.1	*Xanthomonas citri* subsp. *citri* strain gd3	Xccgd3	X	X						X
GCA_000961315.1	*Xanthomonas citri* subsp. *citri* strain jx4	Xccjx4	X	X						X
GCA_000961295.1	*Xanthomonas citri* subsp. *citri* strain jx5	Xccjx5	X	X						X
GCA_000961275.1	*Xanthomonas citri* subsp. *citri* strain mf20	Xccmf20	X	X						X
GCA_000961255.1	*Xanthomonas citri* subsp. *citri* strain MN10	XccMN10	X	X						X
GCA_000961235.1	*Xanthomonas citri* subsp. *citri* strain MN11	XccMN11	X	X						X
GCA_000961215.1	*Xanthomonas citri* subsp. *citri* strain MN12	XccMN12	X	X						X
GCA_000961195.1	*Xanthomonas citri* subsp. *citri* strain NT17	XccNT17	X	X						X
GCA_000961155.1	*Xanthomonas citri* subsp. *citri* strain UI7	XccUI7	X	X						X
GCA_002139975.1	*Xanthomonas citri* subsp. *citri* TX160042	XccTX160042	X	X						
GCA_002139955.1	*Xanthomonas citri* subsp. *citri* TX160149	XccTX160149	X	X						

(Continuação)

Acesso	Nome	Abreviação	Fito-patógeno	Associado a plantas	Usado no teste com 10 genomas	Usado no teste com 20 genomas	Usado no teste com 30 genomas	Usado no teste com 40 genomas	Usado no teste com 50 genomas	Usado no teste com 69 genomas
GCA_002139995.1	*Xanthomonas citri* subsp. *citri* TX160197	XccTX160197	X	X						
GCA_000961175.1	*Xanthomonas citri* subsp. *citri* UI6	XccUI6	X	X						X
GCA_002288565.1	*Xanthomonas citri* subsp. *malvacearum* AR81009	XcimalvAR81009	X	X						
GCA_002288585.1	*Xanthomonas citri* subsp. *malvacearum* MS14003	XcimalvMS14003	X	X						
GCA_001719145.1	*Xanthomonas citri* subsp. *malvacearum* MSCT	XcimalvMSCT	X	X						
GCA_002224525.1	*Xanthomonas citri* subsp. *malvacearum* XcmH1005	XcimalvXcmH1005	X	X						
GCA_002224545.1	*Xanthomonas citri* subsp. *malvacearum* XcmN1003	XcimalvXcmN1003	X	X						
GCA_001908795.1	*Xanthomonas euvesicatoria* LMG930	XeuvesicLMG930	X	X						
GCA_001705565.1	*Xanthomonas fragariae*	Xfragariae	X	X						
GCA_900183985.1	*Xanthomonas fragariae* NBC2815	XfragariaeNBC2815	X	X						
GCA_900183995.1	*Xanthomonas fragariae* PD5205	XfragariaePD5205	X	X						
GCA_900183975.1	*Xanthomonas fragariae* PD885	XfragariaePD885	X	X						
GCA_001610915.1	*Xanthomonas fuscans* subsp. *aurantifolii* 1566	Xfaurantifolii1566	X	X						
GCA_001610795.1	*Xanthomonas fuscans* subsp. *aurantifolii* FDC 1559	XfaurantifoliiFDC1559	X	X						
GCA_001610815.1	*Xanthomonas fuscans* subsp. *aurantifolii* FDC 1609	XfaurantifoliiFDC1609	X	X						
GCA_000969685.1	*Xanthomonas fuscans* subsp. *fuscans* str. 4834-R, chromosome	Xff4834-R	X	X						X
GCA_001908775.1	*Xanthomonas gardneri* ICMP7383	XgardneriICMP7383	X	X						
GCA_001908755.1	*Xanthomonas gardneri* JS749-3	XgardneriJS749-3	X	X						
GCA_002285515.1	*Xanthomonas hortorum* B07-007	XhortorumB07-007	X	X						
GCA_001466505.1	*Xanthomonas oryzae* pv. *oryzae* AXO1947	XooAXO1947	X	X						

(Continuação)

Acesso	Nome	Abreviação	Fito-patógeno	Associado a plantas	Usado no teste com 10 genomas	Usado no teste com 20 genomas	Usado no teste com 30 genomas	Usado no teste com 40 genomas	Usado no teste com 50 genomas	Usado no teste com 69 genomas
GCA_000007385.1	*Xanthomonas oryzae* pv. *oryzae* KACC 10331	XooKACC10331	X	X						
GCA_000010025.1	*Xanthomonas oryzae* pv. *oryzae* MAFF 311018 DNA	XooMAFF311018DNA	X	X						
GCA_002850135.1	*Xanthomonas oryzae* pv. *oryzae* MAI106	XooMAI106	X	X						
GCA_002850155.1	*Xanthomonas oryzae* pv. *oryzae* MAI129	XooMAI129	X	X						
GCA_002850175.1	*Xanthomonas oryzae* pv. *oryzae* MAI134	XooMAI134	X	X						
GCA_002850095.1	*Xanthomonas oryzae* pv. *oryzae* MAI145	XooMAI145	X	X						
GCA_002850115.1	*Xanthomonas oryzae* pv. *oryzae* MAI68	XooMAI68	X	X						
GCA_002850075.1	*Xanthomonas oryzae* pv. *oryzae* MAI73	XooMAI73	X	X						
GCA_002850195.1	*Xanthomonas oryzae* pv. *oryzae* MAI95	XooMAI95	X	X						
GCA_002850215.1	*Xanthomonas oryzae* pv. *oryzae* MAI99	XooMAI99	X	X						
GCA_001746615.1	*Xanthomonas oryzae* pv. *oryzae* PXO145	XooPXO145	X	X						
GCA_001746635.1	*Xanthomonas oryzae* pv. *oryzae* PXO211	XooPXO211	X	X						
GCA_001746655.1	*Xanthomonas oryzae* pv. *oryzae* PXO236	XooPXO236	X	X						
GCA_001746675.1	*Xanthomonas oryzae* pv. *oryzae* PXO282	XooPXO282	X	X						
GCA_001746695.1	*Xanthomonas oryzae* pv. *oryzae* PXO524	XooPXO524	X	X						
GCA_001746715.1	*Xanthomonas oryzae* pv. *oryzae* PXO563	XooPXO563	X	X						
GCA_001746735.1	*Xanthomonas oryzae* pv. *oryzae* PXO602	XooPXO602	X	X						
GCA_001746595.1	*Xanthomonas oryzae* pv. *oryzae* PXO71	XooPXO71	X	X						
GCA_001518895.1	*Xanthomonas oryzae* pv. *oryzae* PXO83	XooPXO83	X	X						
GCA_000948075.1	*Xanthomonas oryzae* pv. *oryzae* PXO86	XooPXO86	X	X						X

(Continuação)

Acesso	Nome	Abreviação	Fito-patógeno	Associado a plantas	Usado no teste com 10 genomas	Usado no teste com 20 genomas	Usado no teste com 30 genomas	Usado no teste com 40 genomas	Usado no teste com 50 genomas	Usado no teste com 69 genomas
GCA_000019585.2	*Xanthomonas oryzae* pv. *oryzae* PXO99A	XooPXO99A	X	X						X
GCA_002023005.1	*Xanthomonas oryzae* pv. *oryzae* XF89b	XooXF89b	X	X						
GCA_000168315.3	*Xanthomonas oryzae* pv. *oryzicola* BLS256	XooryzicolaBLS256	X	X						X
GCA_001042745.1	*Xanthomonas oryzae* pv. *oryzicola* strain B8-12	XooryzicolaB8-12	X	X						X
GCA_001042775.1	*Xanthomonas oryzae* pv. *oryzicola* strain BLS279	XooryzicolaBLS279	X	X						X
GCA_001042795.1	*Xanthomonas oryzae* pv. *oryzicola* strain BXOR1	XooryzicolaBXOR1	X	X						X
GCA_001042735.1	*Xanthomonas oryzae* pv. *oryzicola* strain CFBP2286	XooryzicolaCFBP2286	X	X						X
GCA_001042815.1	*Xanthomonas oryzae* pv. *oryzicola* strain CFBP7331	XooryzicolaCFBP7331	X	X						X
GCA_001042835.1	*Xanthomonas oryzae* pv. *oryzicola* strain CFBP7341	XooryzicolaCFBP7341	X	X						X
GCA_000940825.1	*Xanthomonas oryzae* pv. *oryzicola* strain CFBP7342	XooryzicolaCFBP7342	X	X						X
GCA_001042855.1	*Xanthomonas oryzae* pv. *oryzicola* strain L8	XooryzicolaL8	X	X						X
GCA_001042875.1	*Xanthomonas oryzae* pv. *oryzicola* strain RS105	XooryzicolaRS105	X	X						X
GCA_001021915.1	*Xanthomonas oryzae* pv. *oryzicola* strain YM15	XooryzicolaYM15	X	X						X
GCA_000192045.3	*Xanthomonas perforans* 91-118	Xperforans91-11891-118	X	X						
GCA_001908855.1	*Xanthomonas perforans* LH3	XperforansLH3	X	X						
GCA_002759095.1	*Xanthomonas phaseoli* pv. *phaseoli* CFBP412	XppCFBP412	X	X						
GCA_002759115.1	*Xanthomonas phaseoli* pv. *phaseoli* CFBP6164	XppCFBP6164	X	X						
GCA_002759135.1	*Xanthomonas phaseoli* pv. *phaseoli* CFBP6546R	XppCFBP6546R	X	X						
GCA_002759155.1	*Xanthomonas phaseoli* pv. *phaseoli* CFBP6982	XppCFBP6982	X	X						
GCA_000815185.1	*Xanthomonas sacchari* strain R1	XsaccharíR1		X						X

(Continuação)

Acesso	Nome	Abreviação	Fito-patógeno	Associado a plantas	Usado no teste com 10 genomas	Usado no teste com 20 genomas	Usado no teste com 30 genomas	Usado no teste com 40 genomas	Usado no teste com 50 genomas	Usado no teste com 69 genomas
GCA_900094325.1	*Xanthomonas translucens* pv. *translucens* DSM 18974	XttDSM18974	X	X						X
GCA_001021935.1	*Xanthomonas translucens* pv. *undulosa* strain Xtu 4699	XtuXtu4699	X	X						X
GCA_001908725.1	*Xanthomonas vesicatoria* ATCC 35937 LMG911	XvesicATCCLMG911	X	X						
GCA_001908815.1	*Xanthomonas vesicatoria* LM159	XvesicLM159	X	X						
GCA_001456195.1	*Xylella fastidiosa* 3124	Xyf3124	X	X						
GCA_000006725.1	*Xylella fastidiosa* 9a5c	Xyf9a5c	X	X						X
GCA_001456335.3	*Xylella fastidiosa* Fb7	XyfFb7	X	X						
GCA_001456315.1	*Xylella fastidiosa* Hib4	XyfHib4	X	X						
GCA_001456235.1	*Xylella fastidiosa* J1a12	XyfJ1a12	X	X						
GCA_000019325.1	*Xylella fastidiosa* M12	XyfM12	X	X						X
GCA_000019765.1	*Xylella fastidiosa* M23	XyfM23	X	X						X
GCA_000698825.1	*Xylella fastidiosa* MUL0034	XyfMUL0034	X	X						X
GCA_001456295.1	*Xylella fastidiosa* Pr8x	XyfPr8x	X	X						
GCA_000148405.1	*Xylella fastidiosa* subsp. *fastidiosa* GB514	XyffGB514	X	X						X
GCA_002117875.1	*Xylella fastidiosa* subsp. *pauca* De Donno	XyfDeDonno	X	X						
GCA_000698805.1	*Xylella fastidiosa* subsp. sandyi Ann-1	XyfSandyiAnn-1	X	X						X
GCA_000007245.1	*Xylella fastidiosa* Temecula1	XyfTemecula1	X	X						X
GCA_001456275.1	*Xylella fastidiosa* U24D	XyfU24D	X	X						

Fonte: (SANTIAGO et al., 2019)

Anexo B – Distribuição genomas de acordo com os grupos de genomas

B.1 Streptococcus pyogenes

Quadro 6 – Quantidade de genomas de acordo com as anotações doenças, invasividade e padrão para o conjunto de genomas de *Streptococcus pyogenes*

Grupo	Tipo	Quantidade de Genomas
Doenças	Acute Phlegmonous Gastritis	1
	Acute Poststreptococcal Glomerulonephritis	3
	Acute Rheumatic Fever	4
	Bacteremia	4
	Cerebrospinal Fluid Infection	1
	Endometritis	2
	Impetigo	2
	Meningitis	2
	Necrotizing Fasciitis	7
	Perianal Streptococcal Cellulitis	2
	Pharyngitis	9
	Puerperal Sepsis	1
	Scarlet Fever	4
	Skin Infection	1
	Soft Tissue Infection	1
	Streptococcal Toxic Shock Syndrome	4
	Subcutaneous Abscess	1
	Superficial Dermatitis	1
	U	4
	Wound Infection	1
Invasividade	Invasivo	28
	Não-invasivo	25
	Desconhecido	2
Padrão	A-C	25
	D	10
	E	20

Fonte: Caio Santiago, 2019

Quadro 7 – Quantidade de genomas de acordo com o genótipo *emm* para o conjunto de genomas de *Streptococcus pyogenes*

Grupo	Tipo	Quantidade de Genomas
Genótipo *emm*	28	5
	2	1
	1	10
	89	5
	59	2
	66	1
	49	1
	71	1
	82	1
	83	3
	101	1
	44	2
	12	4
	5	1
	23	1
	6	3
	3	4
	53	2
	80	1
	18	1
	14	1
	87	1
	114	1
	4	2

Fonte: Caio Santiago, 2019

B.2 Xanthomonadaceae

Quadro 8 – Quantidade de genomas de acordo com os grupos de genomas anotados para o conjunto de 69 genomas da família *Xanthomonadaceae*

Grupo	Tipo	Quantidade de Genomas
Fito-associados	Sim	58
	Não	11
Fito-patogênicos	Sim	57
	Não	12

Fonte: Caio Santiago, 2019

Anexo C – Quantidade de famílias exclusivas encontradas de acordo com cada um dos grupos de genomas

C.1 Genótipo emm *do estudo de caso dos* Streptococcus pyogenes

Tabela 6 – Quantidade de famílias exclusivas encontradas no conjunto de genomas de *Streptococcus pyogenes*, considerando apenas a anotação dos grupos de genomas do genótipo *emm*.

Genótipo *emm*	Quantidade de famílias exclusivas	Quantidade de genomas
1	14	10
2	52	1
3	14	4
4	20	2
5	40	1
6	16	3
12	6	4
14	35	1
18	30	1
23	52	1
28	19	5
44	17	2
49	61	1
53	3	2
59	12	2
66	39	1
71	38	1
80	41	1
82	30	1
83	2	3
87	40	1
89	14	5
101	22	1
114	56	1

Fonte: Caio Santiago, 2019

C.2 Doenças do estudo de caso dos Streptococcus pyogenes

Tabela 7 – Quantidade de famílias exclusivas encontradas no conjunto de genomas de *Streptococcus pyogenes*, considerando apenas a anotação dos grupos de genomas das doenças causadas por *Streptococcus pyogenes*.

Doenças	Quantidade de famílias exclusivas	Quantidade de genomas
Acute Phlegmonous Gastritis	20	1
Acute Poststreptococcal Glomerulonephritis	0	3
Acute Rheumatic Fever	0	4
Bacteremia	0	4
Cerebrospinal Fluid Infection	15	1
Endometritis	0	2
Impetigo	2	2
Meningitis	4	2
Necrotizing Fasciitis	0	7
Perianal Streptococcal Cellulitis	0	2
Pharyngitis	0	8
Puerperal Sepsis	20	1
Scarlet Fever	0	4
Skin Infection	35	1
Soft Tissue Infection	20	1
Streptococcal Toxic Shock Syndrome	0	5
Subcutaneous Abscess	39	1
Superficial Dermatitis	48	1
Unknown	0	4
Wound Infection	17	1

Fonte: Caio Santiago, 2019

C.3 Ferramentas e parâmetros utilizados

C.3.1 Pré-processamento

- **Anotação de genomas:**
  - Ferramenta: Patric Annotation Service baseada em RASTk (WATTAM et al., 2017).
  - Parâmetros: Valores iniciais.
- **Comparação de sequências codificantes:**
  - Ferramenta: Blast.
  - Parâmetros: E-value mínimo de 10^{-10}.
- **Clusterização de sequências (homólogos):**

  - Ferramenta: GTACG.
  - Parâmetros: E-value mínimo de 10^{-10} e comprimento mínimo do alinhamento variável em relação ao problema tratado. No caso do conjunto de *S. pyogenes* o comprimento mínimo dos alinhamentos foi de 42% e para o conjunto de *Xanthomonadaceae* foi de 45%.

- **Alinhamento das famílias de homólogos**
  - Ferramenta: Clustal Omega.
  - Parâmetros: Valores iniciais.
- **Filogenia das famílias de homólogos**
  - Ferramenta: FastTree.
  - Parâmetros: Valores iniciais.
- **Clusterização de sequências (ortólogos):**
  - Ferramenta: GTACG inspirado no método desenvolvido por Ding, Baumdicker e Neher (2018).
  - Parâmetros: remoção de ramos filogenéticos maiores que 0,4.
    A partir da filogenia de cada uma das famílias, os ramos que são maiores que 0,4 são removidos, subdividindo assim as famílias em dois ou mais subgrupos.
- **Identificação de proteínas multidomínio**
  - Ferramenta: GTACG.
  - Parâmetros: Para ser considerada multidomínio a proteína precisa ter um coeficiente de agrupamento menor que a média de suas homólogas, além disso, o alinhamento entre a proteína multidomínio e as de domínio único deve apresentar mais do que 30% de diferença e esta diferença precisa ter no mínimo 100 aminoácidos.

C.3.2 Comparação de genomas

Filogenias:

- **Supertree**
  - Ferramenta: Clann (CREEVEY; MCINERNEY, 2005)
  - Parâmetros:
    * Critério: Quartets Fit (QFIT).
    * Heurística de busca: neighbour interchange (NNI).

    - Amostras: 1.
    - Repetições: 1.
    - Máximo de trocas: 1.

- **Consenso**
  - Ferramenta: Phylip consense (FELSENSTEIN, 2005).
  - Parâmetros: Algoritmo Majority Rule Extended.
- **Matriz de distância**
  - Ferramenta: Phylip neighbor (FELSENSTEIN, 2005).
  - Parâmetros: Valores iniciais.
- **Vetor de características**
  - Ferramenta: Phylip pars (FELSENSTEIN, 2005).
  - Parâmetros: Valores iniciais.

Comparação 2D

- Ferramenta: MDSJ: Java Library for Multidimensional Scaling (GROUP, 2009)
- Parâmetros: Algoritmo Classical Scaling