UNIVERSIDADE FEDERAL DO PARANÁ

ANTONIO CAMILO DA SILVA FILHO

Análise comparativa das ferramentas de predição de ilhas genômicas

CURITIBA

2017

ANTONIO CAMILO DA SILVA FILHO

ANÁLISE COMPARATIVA DAS FERRAMENTAS DE PREDIÇÃO DE ILHAS GENÔMICAS

Dissertação apresentada como requisito parcial para obtenção do título de Mestre em Bioinformática pelo Programa de Pós-Graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, Universidade Federal do Paraná, área de concentração em Bioinformática.

Orientador: Profa. Dra. Jeroniza Nunes Marchaukoski

CURITIBA,

2017

Silva Filho, Antonio Camilo da
S586 Análise comparativa das ferramentas de predição de ilhas genômicas / Antonio Camilo da Silva Filho. - Curitiba, 2017.
102 f.; il.: tab.

Orientadora: Jeroniza Nunes Marchaukoski
Dissertação (Mestrado) - Universidade Federal do Paraná, Setor de Educação Profissional e Tecnológica, Curso de Pós-Graduação em Bioinformática.
Inclui Bibliografia.

1. Genomas. 2. Ilhas genômicas. 3. Bioinformática. I. Marchaukoski, Jeroniza Nunes. II. Título. IV. Universidade Federal do Paraná.

CDD 575.12

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PARANÁ
SETOR DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA

Pós-Graduação em Bioinformática WWW.BIOINFO.UFPR.BR
E-mail: bioinfo@ufpr.br Tel: 41 33614906

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em BIOINFORMÁTICA da Universidade Federal do Paraná foram convocados para realizar a arguição da Dissertação de Mestrado de ANTONIO CAMILO DA SILVA FILHO intitulada: Análise comparativa das ferramentas de predição de ilhas genômicas, após terem inquirido o aluno e realizado a avaliação do trabalho, são de parecer pela sua **APROVAÇÃO**.

Curitiba, 11 de Maio de 2017

Dr. Roberto Tadeu Raittz
Presidente/Programa de Pós-graduação em Bioinformática – UFPR

Dr. Dieval Guizelini
Avaliador Interno/Programa de Pós-graduação em Bioinformática – UFPR

Dr. Helisson Faoro
Avaliador Interno/Programa de Pós-graduação em Bioinformática – UFPR/
Instituto Carlos Chagas – FIOCRUZ/PR

Drª. Cyntia Maria Telles Fadel Picheth
Avaliadora Externa/Programa de Pós-graduação em Ciências Farmacêuticas – UFPR

À minha mãe e minha avó, por todo esforço que fizeram para que eu conseguisse estudar.
À minha esposa, pela amizade, amor e companheirismo

AGRADECIMENTOS

É difícil achar palavras para expressar o quanto eu sou grato por todas as pessoas que estiveram ao meu lado no decorrer desses dois anos de pós-graduação. Gostaria de começar agradecendo a minha família, que mesmo estando longe, me apoiaram incondicionalmente. Em especial minha mãe, uma pessoa que sempre vou ter como exemplo, mostrou que nunca devo desistir de meus sonhos. Ela é um dos principais motivos para eu ter cumprido mais esta etapa na minha vida.

Agradeço a Universidade Federal do Paraná e ao Programa de Pós-Graduação em Bioinformática pela oportunidade concedida, juntamente com todos os professores, que passaram o seu conhecimento com a maior atenção e dedicação, e sempre estiveram dispostos a ajudar no que precisei.

A minha querida e amada esposa Camilla, que após todos esses anos estudando juntos, desde a faculdade passando pela especialização e agora o mestrado, nunca deixou de me apoiar um dia sequer, estando sempre ao meu lado nas horas mais difíceis.

Ao Coordenador do Programa de Pós-Graduação em Bioinformática, Prof. Dr. Roberto Tadeu Raittz e ao Prof. Dr. Dieval Guizelini, por todo o conhecimento transmitido e auxílio no desenvolvimento deste trabalho.

A todos os pós-graduandos e professores do Laboratório de Genética Celular e Molecular da Universidade Federal de Belo Horizonte, pela oportunidade de estágio e todo o conhecimento passado.

Aos funcionários da secretaria do Programa de Pós-Graduação em Bioinformática, que com atenção e dedicação auxiliaram quando necessário.

A CAPES pela concessão de minha bolsa de estudos, me incentivando a evoluir intelectualmente.

Por fim, agradeço especialmente a minha orientadora e amiga Profa. Dra. Jeroniza Nunes Marchaukoski, por todo apoio e dedicação ao longo desses dois anos. Sempre me ajudou e auxiliou, passando todas as suas experiências para o melhor desenvolvimento do nosso trabalho.

“How I choose to feel
Is how I am.
I will not lose my faith.
It's an inside job today”

(Eddie Vedder)

RESUMO

Ilhas genômicas são segmentos de DNA em organismos procariotos, que apresentam características que diferem das demais regiões do genoma. As principais características são: conteúdo GC% distinto do genoma; presença de sequências de inserção e repetições diretas; genes associados à mobilidade, tais como integrase e transposase e; genes codificadores de tRNAs que frequentemente flanqueam essas regiões do DNA. A composição genica das ilhas genômicas pode apresentar funções biológicas, nesses casos são classificadas como: ilha de patogenicidade (PAI), ilha metabólica (MIs), ilha de resistência (RIs) e/ou ilha de simbiose (SIs). As ferramentas de predição das ilhas genômicas utilizam as estratégias de análise de comparação de genomas e análise de composição de sequência. A análise comparativa busca identificar regiões distintas em sequências de organismos próximos, enquanto que a análise de composição avalia e relaciona a composição de regiões com as demais regiões do genoma. Este trabalho tem como objetivo avaliar os preditores de ilhas genômicas já desenvolvidos de forma qualitativa e quantitativa a partir de conjuntos de organismos. As ferramentas foram aplicadas em 15 organismos, dos quais *Escherichia coli* CFT073 foi escolhida como controle por ter ilhas genômicas curadas *in vivo* sendo considerada o nosso padrão ouro. Os resultados comparativos com o padrão ouro revelaram que a ferramenta GIPSy obteve o melhor desempenho, cobrindo cerca de 91% da composição e região das ilhas. Seguidas por Alien Hunter, 81%, IslandViewer3, 78%, Predict Bias, 31%, GI Hunter, 17% e Zisland Explorer com 16%. Na análise da comparação das regiões preditas, a ferramenta Alien Hunter apresentou o melhor desempenho. Em segundo, nesse critério, as ferramentas IslandViewer3 e GIPSy tiveram desempenhos semelhantes. As demais ferramentas apresentaram baixo desempenho. As combinações das ferramentas Alien Hunter, GIPSy e IslandViewer3 apresentam melhores resultados na predição das ilhas genômicas nos organismos estudados.

Palavras-chave: Ilhas genômicas; ilhas de patogenicidade; genes de mobilidade; assinatura genômica; fatores de virulência; transferência horizontal de genes.

ABSTRACT

Genomic islands are segments of DNA in prokaryotic organisms that have characteristics that differ from other regions of the genome. The main characteristics are: GC% content; Insertion sequences and direct repeats; Genes associated with mobility, such as integrase and transferase; tRNAs that frequently flank these regions. The genetic composition of the genomic islands may have biological functions, in which they are classified as: pathogenic island (PAI), metabolic island (MIs), resistance islands (RIs) and / or symbiosis island (SIs). Prediction tools for genomic islands use strategies of genomic comparison analysis and sequence composition analysis. The comparative analysis search to identify distinct regions in sequences of nearby organisms, whereas the composition analysis evaluates and relates the composition of regions with the other regions of the genome. This work aims to evaluate the predictors of genomic islands already developed in a qualitative and quantitative way from sets of organisms. The tools were tested in 15 organisms, of which Escherichia coli CFT073 was chosen as control for having genomic islands already cured *in vivo* being considered our gold standard. The comparative results with the gold standard revealed that the GIPSy tools obtained the best performance, covering about 91% of the composition and region of the islands. Followed by Alien Hunter, 81%, IslandViewer3, 78%, Predict Bias, 31%, GI Hunter, 17% and Zisland Explorer with 16%. In the analysis of the intersection of the predicted regions, the tool Alien Hunter presented the best performance. Second, IslandViewer3 and GIPSy tools performed similarly. The other tools presented low performance. The combination of the tools Alien Hunter, GIPSy and IslandViewer3 present better results in the prediction of the genomic islands in the organisms studied.

Keywords: Genomic islands; Pathogenic islands; Mobility genes; Genomic signature; Virulence Factors; Horizontal gene transfer.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE SIGLAS

BLAST	- Acrônimo para ferramenta de busca de alinhamento local de sequências (*Basic Local Alignment Search Tool*, em inglês).
CARD	- Acrônimo para banco de dados de resistência a antibióticos (*Comprehensive Antibiotic Resistance Database*, em inglês).
CDS	- Sequência de codificação.
CGH	- Hipóteses do *core* genoma.
COG	- Acrônimo para banco de dados de agrupamento de proteínas ortólogos (*Clusters of Orthologous Groups*, em inglês).
CV	- Vetor de composição.
CUTG	- Banco de dados de uso de códons (*Codon Usage Database*, em inglês)
DNA	- Ácido desoxirribonucleico.
DR	- Repetições diretas.
GIs	- Ilhas genômicas.
G+C	- Guanina e citosina.
GIV	- Visualizador de Ilhas genômicas (*Genomic Island Visualization*, em inglês).
GIPSy	- Nome do programa e acrônimo para predição de ilhas genômicas (*Genomic Island Prediction Software*, em inglês).
HFR	- Alta frequência de recombinação.
HGT	- Transferência horizontal de genes.
HMM	- Modelo oculto de Markov.
HTML	- Linguagem de marcação de hipertexto.
ICEs	- Elementos integrativos e conjugativos.
IS	- Sequência de inserção.
IVOM	- Motivos de ordem variados e interpolados.
KB	- Quilo byte
MIs	- Ilhas de metabolismo
MGEs	- Elementos genéticos moveis
MvirDB	- Nome do banco de dados de toxinas (*Microbial Database of Protein Toxins*, em inglês).
MVS	- Segmento múltiplo de Viterbi.

NCBI	- *National Center for Biotechnology Information*.
NODMUTDB	- Banco de dados para genes envolvidos em simbiose (*Nodulation Mutant Database*, em inglês).
ORF	- Fase de leitura aberta, acrônimo para Open Read Frame, em inglês.
PAIs	- Ilhas de patogenicidade.
PCR	- Reação da cadeia de polimerase.
PAIDB	- Banco de dados de ilhas de patogenicidade (*Pathogenicity Island Database*, em inglês).
PATRIC	- *Pathosystems Resource Integration*.
PIPs	- Programa para predição de ilhas de patogenicidade (*Pathogenicity Island Prediction* Software, em inglês).
PHP	- Linguagem computacional interpretada livre
PFAM	- Banco de dados de famílias de proteínas
Ris	- Ilhas de resistência
RNA	- Ácido ribonucleico
SD	- Desvio padrão.
SIs	- Ilhas de simbiose.
SQL	- Linguagem de consulta estruturada.
TRA	- *Operon* de transferência.
tRNA	- Ácido ribonucleico de transferência (prefiro transportador).
UFPR	- Universidade Federal do Paraná.
URL	- Endereço de rede de páginas da internet.
VICTORS	- *Virulence Factors*
VFDB	- *Virulence Factor Database*.
WEKA	-*Waikato Environment for Knowledge Analysis.*

SUMÁRIO

1 INTRODUÇÃO

Existe uma abundância de espécies de bactérias na biosfera, e parte delas possui um papel significativo no desenvolvimento das doenças infecciosas e na taxa de mortalidade dos seres humanos. A compreensão dos mecanismos de virulência das bactérias é um ponto crucial para o entendimento e possível desenvolvimento de alternativas para prevenção e/ou tratamento das doenças infecciosas de origem bacteriana. As bactérias possuem a capacidade de adaptação em diversos ambientes, o que lhes assegurava vantagens para a sobrevivência, facilitando à evolução em situações extremas, a partir da aquisição de material genético de outros organismos. Um dos principais fatores de adaptação e evolução das bactérias é proveniente das Ilhas Genômicas (GIs) (WILSON, 2012).

As GIs são agrupamentos de genes resultantes de transferência horizontal de genes (HGT). Essas ilhas apresentam regiões com características que diferem do restante do genoma, como o seu conteúdo de guanina e citosina (G+C), na frequência uso de códons e na composição dinucleotideos. As GIs são frequentemente flanqueadas por genes de tRNA, repetições diretas (DR) e elementos móveis. Entre outros, os elementos móveis mais frequentemente encontrados nas GIs são as integrases e transposases. As integrases são enzimas responsáveis pela integração, recombinação e excisão da ilha, e as transposases enzimas associadas com a mobilização (HACKER et al., 2001).

Uma importante característica das GIs é a vantagens que elas fornecem aos microrganismos, codificando funções acessórias voltadas para resistência a antibiótico e propriedades relacionadas ao metabolismo, simbiose e patogenicidade (HACKER et al., 2001).

A HGT é um mecanismo que possibilita variação genética das bactérias, que não pode ser obtida através de mutações (LANGILLE; HSIAO; BRINKMAN, 2010). Este é um dos processos mais importantes para gerar diversidade e facilitar a propagação de genes nas bactérias, pelo fato do organismo receber um material já preparado e melhorado, aumentando as suas chances de adaptação (WILSON, 2012).

A capacidade de as bactérias transmitirem característica relacionada com patogenicidade ou resistência aos antibióticos é um dos fatores mais estudados e associados as GIs. A grande disseminação da capacidade de resistência aos

antibióticos é um dos grandes problemas que a saúde enfrenta na atualidade, deixando em dúvida o sucesso no tratamento das doenças infecciosas. Essa mudança nas populações bacterianas aumentando seu nível de resistência a diversos antibióticos mostra como em poucas décadas as bactérias tem conseguido se adaptar e evoluir rapidamente. Cada vez mais as GIs estão sendo associadas ao aumento e distribuição de funções de virulência e resistência aos antibióticos devido ao seu importante papel na evolução dos genomas bacterianos (JUHAS et al., 2009).

Com o constante aperfeiçoamento de técnicas de sequenciamento e o progresso da Bioinformática, várias metodologias para predições de GIs vem sendo desenvolvidas. A escolha do método de predição para detecção das ilhas genômicas poderá variar de acordo com o genoma a ser estudado. Para organismos estreitamente relacionados pode-se utilizar ferramentas com abordagens comparativas, já aqeueles sem um organismo de referência são analisados a partir de técnicas de composição de sequência. Com os resultados dessas predições é possível obter um conhecimento maior sobre as ilhas genômicas permitindo observar seu comportamento nos genomas e até mesmo a sua evolução (LANGILLE; HSIAO; BRINKMAN, 2010).

Mesmo com o grande número de ferramentas de predição disponibilizadas pela comunidade cientifica a precisão dos resultados ainda deixa a desejar. A utilização de apenas uma metodologia pode não ser suficiente para obter resultados satisfatórios. O emprego de técnicas em conjunto é uma boa estratégia para tentar abranger as lacunas ainda existentes nas predições de ilhas genômicas de acordo com (LU; LEONG, 2016).

Avaliar qualitativamente e quantitativamente as ferramentas de predição de ilhas genômicas, utilizando um conjunto de organismos e ilhas conhecidas, permite uma maior compreensão dos métodos de predição, a precisão e sensibilidade desses métodos, bem como uma maior compreensão do comportamento das ilhas nos diferentes organismos e do processo de adaptação e evolução genômica. As ferramentas para predição de ilhas genômicas avaliadas nesse trabalho foram: Alien Hunter (VERNIKOS; PARKHILL, 2006), GI Hunter (CHE; WANG; FAZEKAS, 2014), GIPSy (SOARES et al., 2016), IslandViewer3 (DHILLON et al., 2015), Zisland Explorer (WEI et al., 2016) e Predict Bias (PUNDHIR; VIJAYVARGIYA; KUMA, 2008).

1.1 OBJETIVO GERAL

Avaliar qualitativamente e quantitativamente ferramentas de predição de ilhas genômicas já desenvolvidas, utilizando como referência um organismo com ilhas conhecidas curadas *in vitro*, e um grupo de organismos para confrontar os resultados entre os preditores.

1.2 OBJETIVOS ESPECÍFICOS

Os objetivos específicos desta dissertação são:

- Determinar a ferramenta de predição com maior taxa de acerto, em relação as 16 ilhas curadas *in vitro* da bactéria *Escherichia coli* CFT073.
- Comparar as predições de todas as ferramentas sobre os organismos do grupo de testes, relacionando seus resultados similares e únicos para determinar a ferramenta com maior cobertura.
- Avaliar as melhores metodologias e abordagens e predições das ferramentas, estabelecendo estratégias para pesquisa de Ilhas Genômicas.

2 REVISÃO BIBLIOGRÁFICA

2.1 ILHAS GENÔMICAS E SUA POSSÍVEL ORIGEM

Genomas bacterianos têm evoluído e se adaptado ao longo do tempo devido a uma variedade de processos, envolvendo mutações, rearranjos genéticos ou transferência horizontal de genes (HGT). Essa evolução foi capaz de ser observada em consequência do rápido crescimento de genomas sequenciados. Além dos genes do *core* genoma que codificam funções essenciais, existem outros genes nos genomas bacterianos, como os genes acessórios, possivelemente adquiridos por transferência horizontal de genes (HGT). O processo de HGT confere vantagens para as bactérias e sobrevivência em ambientais desfavoráveis, tornando-as adaptáveis ao meio (SCHMIDT; HENSEL, 2004).

Com a aquisição abundante de genes acessórios derivados da transferência horizontal, surgem regiões atípicas no genoma, chamadas de Ilhas Genômicas (GIs). Essas ilhas possuem um papel importante na evolução, adaptação e diversificação nos genomas bacterianos, portando genes que codificam para proteínas com as mais variadas funções, sendo de grande importância os estudos de seus mecanismos de transferência e origem (JUHAS et al., 2009).

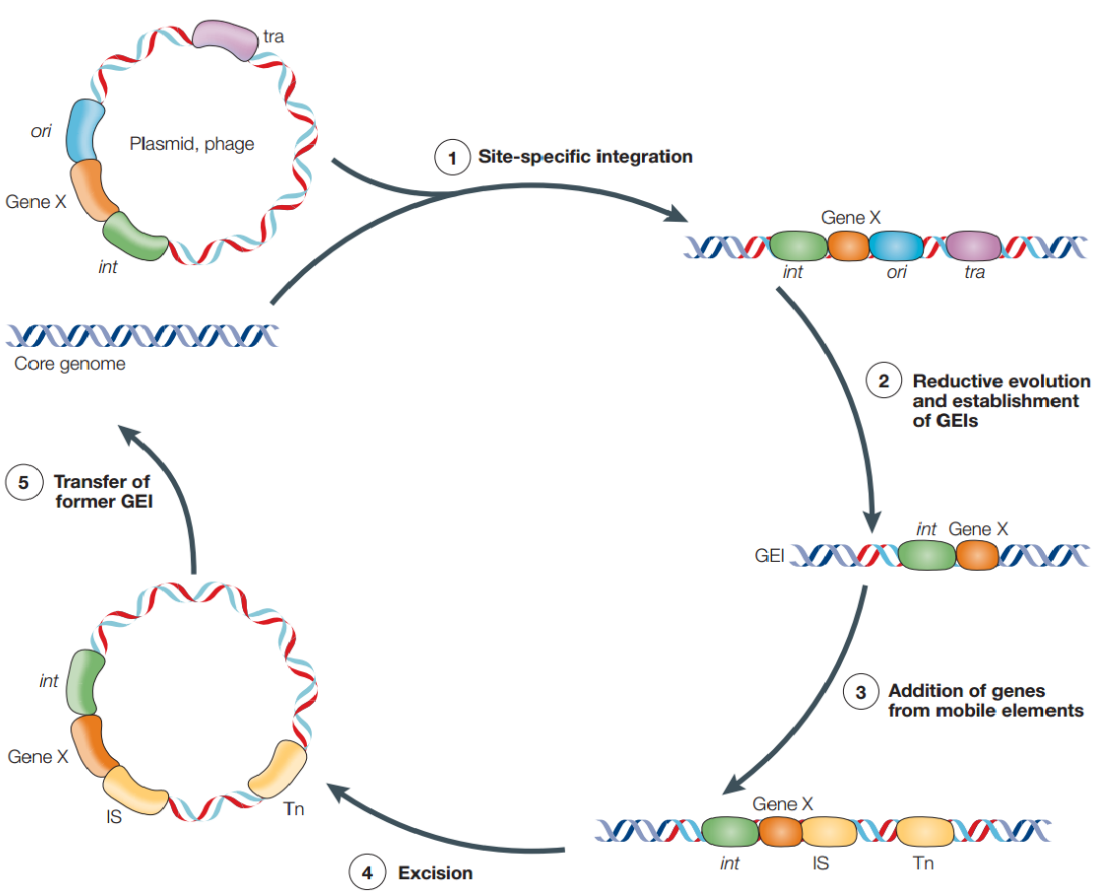
A origem das GIs ainda é incerta. Segundo Bellanger e colaboradores (2014), estudos sugerem que essas ilhas sejam derivadas dos elementos genéticos moveis (MGEs). Nos MGEs estão presentes elementos integrativos e conjugativos que possuem a capacidade de realizar a sua própria integração, transferência e excisão. Após aquisição de material genético através de HGT, os MGEs podem se tornar Gls por rearranjos genéticos, perdas ou ganho de genes.

Os plasmídeos conjugativos ou os fagos de acordo com Burrus e colaboradores (2002) podem ser os principais elementos de origem das GIs pois, após obter uma ligação mais consistente com o cromossomo do hospedeiro, ocorre a perda de genes responsáveis pela sua replicação e auto transferência.

Após a perda dos genes que possibilitam a sua mobilização juntamente com os de replicação, as GIs se tornam imóveis. Entretanto, um dos genes que frequentemente está presente nessas ilhas codifica a Integrase, capaz de exercer função de integração e excisão. GIs podem passar por eventos de recombinação consecutivas, proporcionando ganhos e perdas de informações genéticas, levando

a sua evolução. Por consequência, os MGEs podem ser recuperados, ocasionado a possibilidade de excisão cromossômica da ilha, favorecendo sua transferência para outro organismo (DOBRINDT et al., 2004), O possível ciclo de vida das GIs pode ser representado de acordo com a figura abaixo (figura 1).

FIGURA 1 - POSSÍVEL ORIGEM DAS GI



FONTE: Adaptado de DOBRINDT et al., 2004.
NOTA: Os plasmídeos ou fagos realizam a integração no cromossomo (1) pela perda dos genes responsáveis pela replicação autônoma (ori) e mobilização ou transferência de plasmídeos e bacteriófagos (tra). Com isso se estabelece a GI no cromossomo (2). Devido a evoluções consecutivas, elementos moveis são adquiridos (3), como elementos de inserção (IS) e transposons (Tn). O gene integrase (int) se faz presente na GI possibilitando assim sua excisão do cromossomo (4), contribuindo para sua transferência (5).

Devido a essas propriedades distintas, é possível obter uma melhor compreensão de como essas GIs conseguem se disseminar rapidamente, permitindo que organismos bacterianos evoluam e se adaptem em ambientes desvantajosos, através do seu sistema de conjugação (JUHAS et al., 2007).

2.2 CLASSIFICAÇÃO E COMPOSIÇÃO DAS ILHAS GENÔMICAS

No final da década de 1990, foram descobertos agrupamentos de genes em determinadas estirpes de bactérias que codificaram fatores de virulência e que estavam ausentes em outras da mesma espécie. Esses agrupamentos de genes foram denominados como Ilhas de Patogenicidade (PAI) (HACKER et al., 1990).

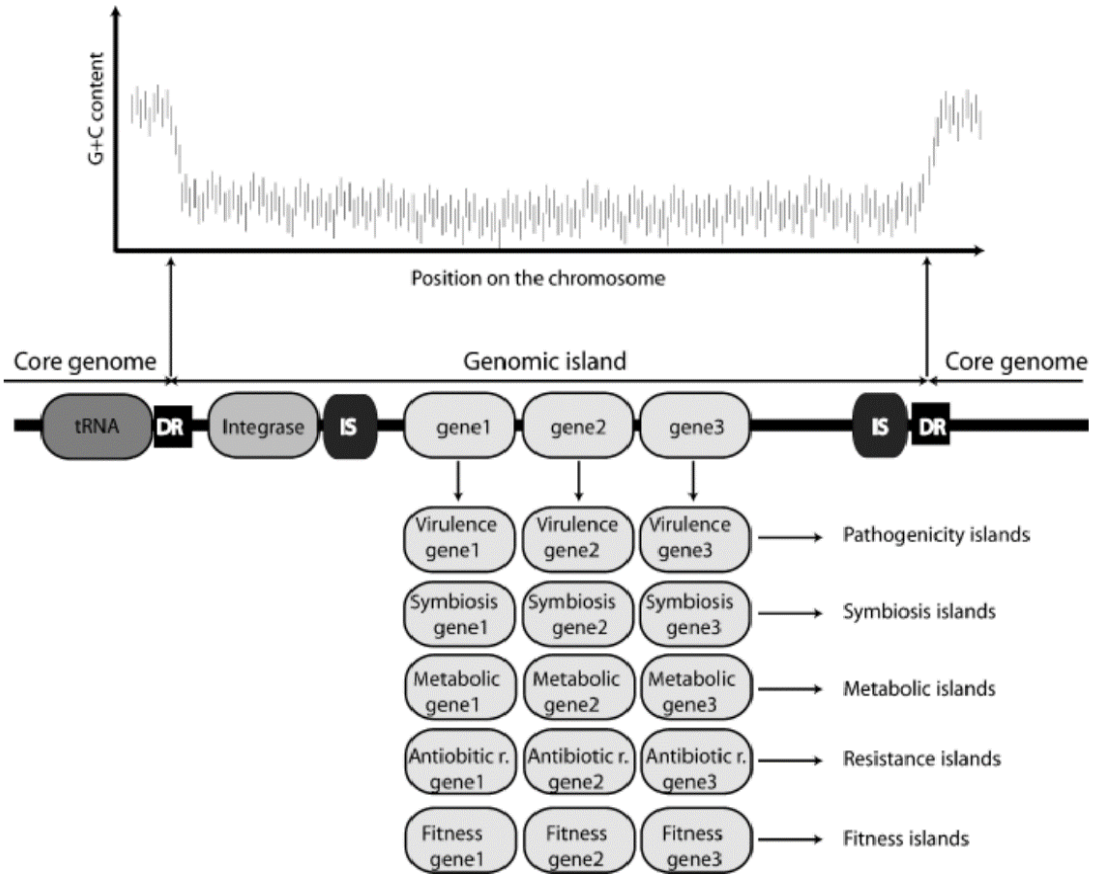
O conceito de PAI foi estabelecido por Hacker e colaboradores (1990). Esses pesquisadores tinham como objetivo estudar a base genética da virulência em estirpes de *Escherichia coli* uropatogênicas. Estudos posteriores destes mesmos autores mostraram que outras classes de GIs poderiam ser estabelecidas dependendo da função biológica que os genes presentes dentro das ilhas exercessem sobre o organismo. Essas classes de GIs foram determinadas como: Ilhas Metabólicas (Mls), contendo genes que codificam proteínas associadas com propriedades metabólicas; ilhas de Resistência (Rls), contendo genes que codificam proteínas associadas com resistência à antibióticos; e ilhas de Simbiose (Sls) (HACKER et al., 1997). Independentemente de sua classe, a maioria das GIs possuem características similares, como:

- Tamanho que varia entre 10 a 200 kb. GIs com número de pares de base abaixo de 10 kb são denominas como Ilhotas Genômicas - *Genomic Islets* (HACKER; KAPER, 2000).
- A composição da sequência difere do restante do genoma, em relação ao conteúdo G+C, frequência uso de códons e dinucleotideos (JUHAS et al., 2009).
- Genes de tRNA se encontram geralmente cercando as GIs e seguidos por sequencias de repetições diretas (DR), podendo atuar como sítios alvos para excisão enzimática (SCHMIDT; HENSEL, 2004).
- Contém genes que codificam integrases ou fatores envolvidos com os processos de conjugação plasmidial ou fagos, relacionados à transferência das ilhas entre os organismos (JUHAS et al., 2009).
- Transposons e elementos de inserção (IS) podem estar presentes, estando relacionados a mobilização e eliminação de material genético (BUCHRIESER et al., 1998; GAL-MOR; FINLAY, 2006).

- De acordo com as funções dos genes presentes, elas podem ser classificadas como, Ilhas de patogenicidade, simbiose, metabólica, resistência a antibióticos e fitness (DOBRINDT et al., 2004; SCHMIDT; HENSEL, 2004).

As características descritas acima são representadas na figura 2:

FIGURA 2. PROPRIEDADES DAS GIS



FONTE: Adaptado de JUHAS et al., 2009.
NOTA: GIs são regiões presentes no genoma que diferem pela sua composição de sequência. Genes de tRNA muitas vezes se encontram a frente dessas ilhas seguidos por sequencias de repetições diretas (DR) contendo genes responsáveis pela mobilidade genética, como integrases, transposases e sequencias de inserção (IS). De acordo as funções biológicas do conjunto de genes que essas ilhas possuem, elas podem ser classificadas como, ilhas de patogenicidade, simbiose, metabólica, resistente a antibióticos e fitness.

As GIs denominadas fitness não dependem somente da presença de genes com funções relacionadas aos fatores de patogenicidade, simbiose, metabolismo ou resistência a antibióticos, mas também do ambiente aonde ela se encontra, proporcionando vários efeitos para o organismo. Isto é, uma mesma ilha em determinadas situações pode exercer funções diferentes (SCHMIDT; HENSEL, 2004).

2.2.1 Ilhas de Patogenicidade (PAI)

Estudos sobre variabilidade genética e evolução de patógenos bacterianos tem grande interesse na área científica. Com a melhor compreensão dos elementos e suas funções que levam esses genomas bacterianos a evoluir rapidamente, as PAIs têm demonstrado a capacidade de facilitar esses processos (DOBRINDT et al., 2004).

Essas ilhas podem estar presentes nos genomas bacterianos de organismos patogênicos, mas se encontram ausentes em organismos não patogênicos da mesma espécie ou estreitamente relacionados. O primeiro relato dessas ilhas nas bactérias ocorreu em decorrência da presença dessas regiões atípicas no DNA cromossômico de *Escherichia coli* (HACKER et al., 1990), no entanto, com o aumento do número de genomas sequenciados e estudos relacionados aos elementos extra cromossômicos, verificou-se que PAIs também podem fazer parte dos genomas de plasmídeo ou bacteriófagos (HACKER; KAPER, 2000).

Devido aos mecanismos de HGT, as bactérias são capazes de ganhar rapidamente funções de virulência proveniente das PAIs. Essas ilhas tem a capacidade de codificar fatores de virulência e outras proteínas acessórias, desempenhando funções, como por exemplo: na aquisição de ferro, produção de toxinas, adesinas e sistemas de secreção (HENSEL, 2004; NOVICK; YORK, 2013).

2.2.2 Características das Ilhas de Patogenicidade

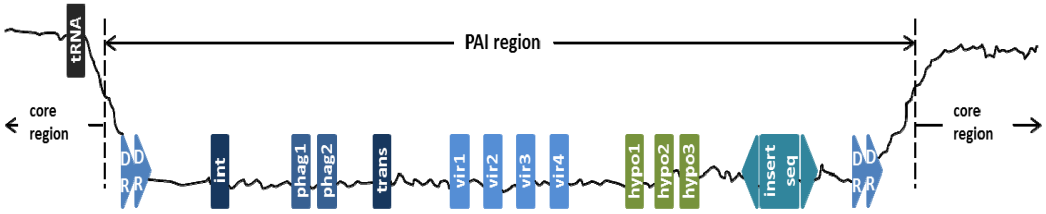
As características das PAIs em comparação com as GIs ou suas subclasses são muito semelhantes, como desvio conteúdo G+C, frequência de códons, presença de repetições diretas, sequencias de inserção e genes relacionados a mobilização. Contudo, há elementos que podem distinguir as PAIs de outras GIs (HACKER; KAPER, 2000).

Os genes associados com virulência estão presentes, como por exemplo as invasinas, que auxiliam as bactérias a invadirem células epiteliais eucarióticas, sistema de captação de ferro, sistema de secreção tipo III e VI que atuam na interação com o hospedeiro para exportação de proteínas, e várias toxinas como, exotoxinas, proteases, lipases e enterotoxinas.

Proteínas hipotéticas se encontram em grande número nas PAIs, se comparado ao *core* genoma. Isto pode ocorrer devido a presença de sequências provenientes de plasmídeos e fagos que ainda não foram sequenciados nessa região, portanto, suas possíveis funções se apresentam indeterminadas (HSIAO et al., 2005). Há presença de grande porcentagem de genes relacionados a fagos, devido ao mecanismo de transdução ser um dos principais meios de transferência de material genético em organismos procariotos (CHE; HSAN; CHEN, 2014).

A transferência das PAIs, ou outras GIs que possuam fatores promovendo vantagens aos microrganismos, possibilita a sua constante adaptação e evolução sendo facilitada pelos mecanismos de HGT (JUHAS et al., 2009). Na figura abaixo (figura 3), é possível visualizar uma ilustração de uma suposta PAI genérica e suas características distintas.

FIGURA 3. EXEMPLO DE UMA PAI E SEUS PRODUTOS



FONTE: Adapatado de CHE; HASAN; CHEN, 2014
NOTA: Região candidata a PAI contendo gene de tRNA, repetições diretas (DR), genes relacionados a mobilização como integrase (int) e transposase (trans), relacionados a fagos (phag1, phag2), genes com fatores de virulência (vir1, vir2, vir3, vir4), proteínas hipotéticas (hypo1, hypo2, hypo3), sequencias de inserção (insert seq).

2.3 TRANSFERÊNCIA HORIZONTAL DE GENES (HGT)

A transferência horizontal, é um dos processos mais importantes para gerar diversidade e facilitar a propagação de genes que proporcionem vantagens para os microrganismos, aumentando suas chances de adaptação e evolução (WILSON, 2012).

A transmissão de material genético entre os microrganismos foi observada pela primeira vez por (TATUM; LEDERBERG, 1947). Com os avanços da tecnologia de sequenciamento genômico nas últimas décadas, foi possível pesquisar mais a fundo sobre os processos de HGT, e sua distribuição no meio bacteriano (BERG; KURLAND, 2002).

Em decorrência do grande volume de genomas sequenciados, pode-se observar indícios de evolução bacteriana por meio de análises das características do genoma, como sua composição genética. Os teores de conteúdo G+C e frequência de códons apresentavam variações, espécies distantes possuíam genes com alta similaridade, estirpes estreitamente relacionadas indicavam alteração de conteúdo e visualização das árvores filogenéticas era inconsistente (KOONIN; MAKAROVA; ARAVIND, 2001).

Lan e Reeves (2000) sugeriram um conceito sobre genomas de espécies bacterianas que dispõem do mesmo conjunto de genes partilhados por todas as estirpes da mesma espécie, definido como “hipótese do *core* genoma” (CGH), e genes auxiliares/acessórios, ajudando na identificação de genes que possam ser originários de HGT.

O principal motivo de aquisição de material genético é devido a necessidade de sobrevivência da bactéria. Suas limitações em ambientes hostis conduzem os organismos a buscarem meios de superar essas dificuldades, sendo os mecanismos de HGT um dos principais fatores para que isso aconteça (HACKER; KAPER, 2000).

2.3.1 Mecanismos de Transferência Horizontal de genes

A transferência de material genético entre organismos pode ocorrer a partir de três processos, denominados: transformação (absorção de DNA livre), transdução (mediada por bacteriófagos), e conjugação (transferência de genes através de plasmídeos ou elementos conjugativos). Em cada caso, as células que compartilham seu material genético são chamadas de “doadoras”, e as que recebem são denominadas “receptoras” (MAIDEN, 1998).

Transformação é o mecanismo que possibilita as bactérias absorverem e integrar ao genoma o material genético que se encontram no meio extracelular (GRIFFITH, 1981). Nos domínios de Bactéria e Archea, este processo de transferência demonstrou estar presente somente em bactérias competentes, capazes de receber DNA exógeno (LORENZ; WACKERNAGEL, 1994). Qualquer célula que seja capaz de absorver esse material, é denominada “competentes”, essas células normalmente são induzidas pelo ambiente em que se encontram. A partir da lise celular, a célula libera todo seu material genético no meio extracelular, esse material é fragmentando por conta do seu grande tamanho. Para que esse material

seja absorvido, uma serie de proteínas ajudam nesse processo para que a membrana da célula tenha condições de realizar a passagem desse material (CHEN; DUBNAU, 2004).

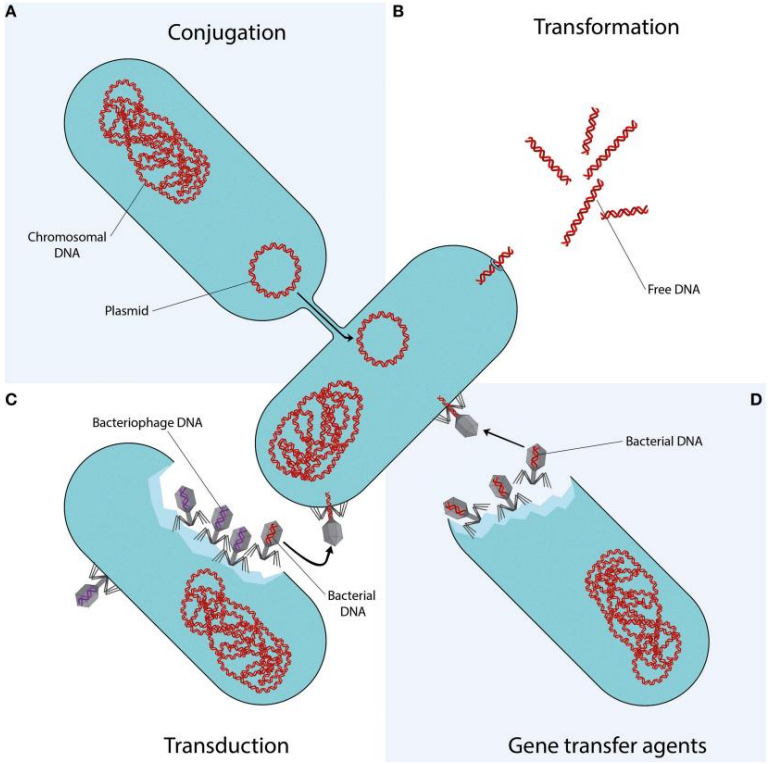
Transdução ocorre através dos bacteriófagos também chamados de fagos ou agentes de transferência de genes (LWOFF, 1953). Os bacteriófagos (vírus que infectam organismos procariotos) após adentrarem na bactéria, tem a capacidade de adquirir segmentos de material genético da célula hospedeira em seu capsídeo e, após a lise celular, é capaz de inseri-lo em outro hospedeiro (FROST et al., 2005).

A sua replicação pode ocorrer a partir de dois ciclos, lítico e lisogênico. No ciclo lítico, após a transmissão do material genético dos bacteriófagos para dentro da célula, este é incorporado pelo genoma hospedeiro ocasionando replicação descontrolada desses vírus dentro da célula, levando a lise celular. Já no ciclo lisogênico, acontece a integração do material genético do vírus ao DNA bacteriano, utilizando todas as propriedades da célula afim de se reproduzir (MAURICE, 2013).

A conjugação é o processo em que as bactérias a partir do contato da célula doadora com a receptora adquirem material genético (LEDERBERG; TATUM, 1946). Para que este processo aconteça, a célula doadora do material genético deve dispor de um plasmídeo conjugativo ou plasmídeo F, podendo estar integrado ou não no cromossomo da bactéria. Se houver a ocorrência desse plasmídeo de forma integrada, a bactéria será denominada então de HFr (alta frequência de recombinação), plasmídeo F+ (doadora), e as bactérias desprovidas plasmídeos F- (receptora) (BABIC et al., 2008). O processo de conjugação está relacionado com a presença de genes que se fazem presentes no operon de transferência (tra), responsáveis pela síntese da pilus F, possibilitando o reconhecimento e contato entre as células, como a transferência do material genético (GROHMANN; MUTH; ESPINOSA, 2003).

Na figura 4, representa os três mecanismos resumidamente.

FIGURA 4. MECANISMOS DE TRANSFERÊNCIA HORIZONTAL DE GENES (HGT)



FONTE: VON WINTERSDORFF et al., 2016
NOTA: Os quadros (A) conjugação, (B) transformação, (C) transdução, representa cada mecanismo de HGT e o quadro (D) designa os bacteriófagos *"Gene transfer agents"*.

2.3.2 Elementos Genéticos Móveis

Os elementos genéticos moveis (MGEs), além de possuírem a capacidade de mobilização, normalmente transportam genes que contribuem para facilitar a sua transferência (FROST et al., 2005).

Há duas classificações para os MGEs, podendo ser: elementos com mobilidade intracelular e; elementos com mobilidade intercelular. Os elementos com mobilidade intercelur são mediados, em organismos procariotos, pelos mecanismos de transformação, transdução e conjugação (RANKIN; ROCHA; BROWN, 2010). Alguns dos principais MGEs presentes nas GIs são conceituados a seguir, estão representados na figura 5:

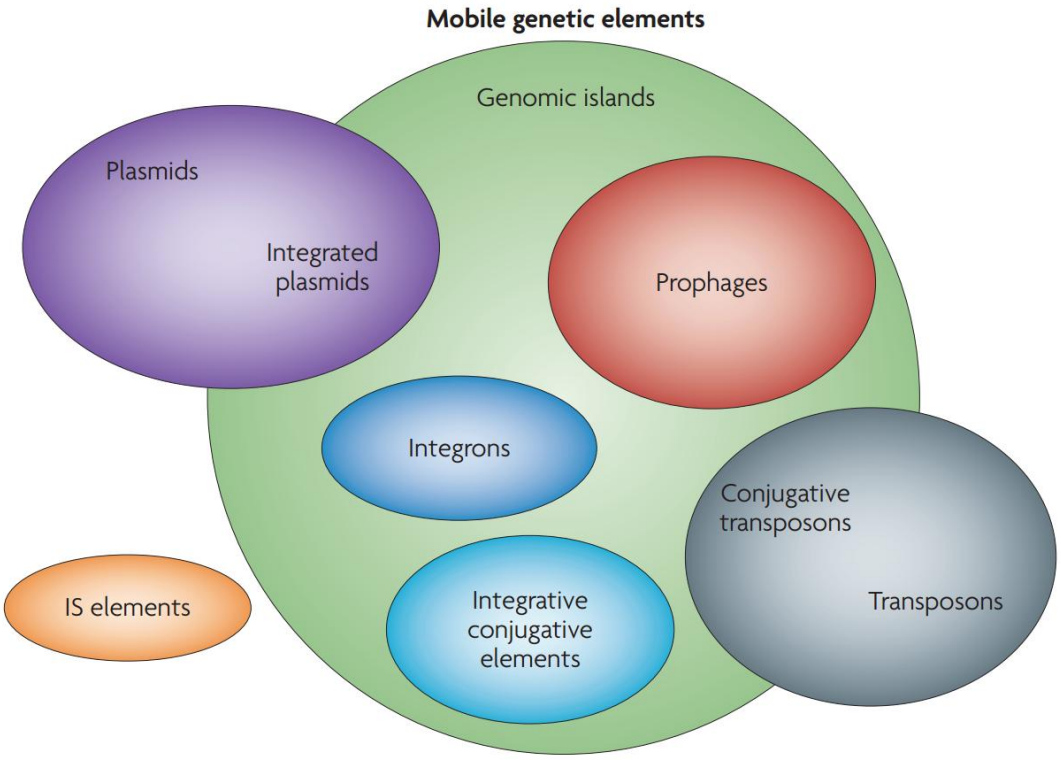
- Plasmídeo / Plasmídeo Integrativo: DNA extracromossômico circular tendo a capacidade de auto replicar-se. Normalmente possuem genes que podem trazer algum benefício para a bactéria aumentando suas

chances de adaptação e sobrevivência. Plasmídeo conjugativo possui a capacidade de realizar mecanismo de conjugação (THOMAS, 2000).

- Bacteriófagos: Correspondem a vírus que infectam apenas células de organismos procariotos, e podem realizar o ciclo lítico, no qual a maquinaria de biossíntese bacteriana é utilizada para gerar partículas virais e culmina com a lise da bactéria ou o ciclo lisogênico no qual o DNA viral é incorporado ao cromossomo da bactéria (CANCHAYA et al., 2003).
- Integrons: São elementos genéticos que tem a capacidade de codificar integrase, e podem conter genes de resistência antibióticos. Sua capacidade de mobilização é limitada, não podendo realizar esta ação por conta própria. São comumente encontrado em plasmídeos e transposons, que ajudam na sua deslocação (CAMBRAY; GUEROUT; MAZEL, 2010).
- Elementos integrativos e conjugativos (ICEs): Designam grupos de MGEs que possuem características em comuns, como os plasmídeos e bacteriófagos. São auto transmissíveis e podem estar integrados no genoma sendo difundidos durante a replicação cromossômica, divisão celular e mecanismos de conjugação (BURRUS; WALDOR, 2004).
- Sequência de inserção (IS): São semelhantes aos transposon, e só contém o gene que codifica integrase para catalisar a sua transposição. Esses elementos podem alterar a expressão de genes próximos, interromper sequencias codificantes ou regiões regulatórias (SIGUIER et al., 2006).
- Transposons / Transposons conjugativos: Tem a característica de se mover nos genomas. Podem apresentar dois mecanismos de transposição. O primeiro tem capacidade de copiar o material genético de uma determinada região do genoma e inseri-lo em outra sem que ocorra a eliminação do local original. Segunda elimina o seu local de transferência inicial e transloca o material para outra região através da enzima transposase. Diferente do transposon conjugativo, que pode se locomover entre genomas por meio do mecanismo de conjugação e

integrando-se no cromossomo da célula receptora (KAPER; HACKER, 1999).

FIGURA 5. PRINCIPAIS MGEs PRESENTES EM GIs



FONTE: Adaptado de LANGILLE; HSIAO; BRINKMAN, 2010.

2.4 GIS NA EVOLUÇÃO DE GENOMAS BACTERIANOS

A evolução genômica pode ocorrer através de mudanças no conteúdo genético de um organismo ao longo do tempo, a partir de mutações, conversões de genes, rearranjos, deleções e inserções de genes, causando grande alterações nos genomas, podendo alterar drasticamente o estilo de vida de uma bactéria (GROISMAN; OCHMAN, 1996).

Cada vez mais, pesquisas estão sendo realizadas para compreender melhor os mecanismos que levam os genomas bacterianos a evoluir e se adaptar continuadamente. Muitos indícios trazem os elementos genéticos moveis como um dos principais fatores para que isso ocorra, sendo as GIs responsáveis por parte da diversidade genética dos microrganismos (DOBRINDT et al., 2004).

Com a grande quantidade de organismos sendo sequenciados e a evolução continua de técnicas de análises de sequências genéticas, foi possível perceber que

as GIs são um mosaico de genes formados através de mecanismos de HGT, que podem trazer benefícios e vantagens para os microrganismos, ajudando na sua adaptação e sobrevivência, ultrapassando as barreiras de gênero e espécie (HACKER et al., 2001).

2.5 MÉTODOS DE PREDIÇÕES GIS

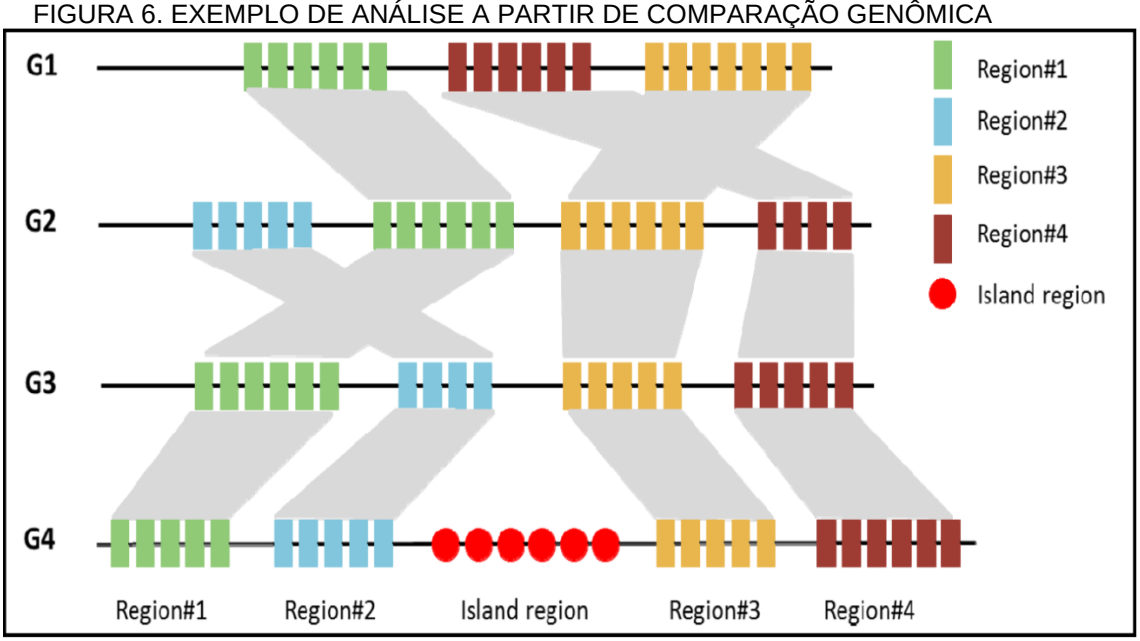
Em 1990, Hacker e colaboradores identificaram a primeira PAI por meio de experimentos em biologia molecular. Com o constante avanço das tecnologias foi possível desenvolver ferramentas computacionais para auxiliar nas predições (SOARES; OLIVEIRA; JAISWAL, 2016). Análises para identificação de possíveis GIs canditadas nos organismos podem ser realizados por meios experimentais e computacionais. Estudos *in silico* (simulação computacional) são mais utilizados, devido ao grande tempo necessário para reproduzir experimentos *in vitro* e também seu alto custo. Os principais métodos que as ferramentas de predição utilizam se dividem em dois grupos: Análise de genômica comparativa, com o objetivo de identificar regiões que se encontram ausentes em organismos relativamente próximos (múltiplos genomas) e análise de composição de sequência do organismo (único genoma) (LU; LEONG, 2016).

2.5.1 Análise genômica comparativa

Este método de análise é utilizado quando genomas de referência se fazem presentes, podendo diminuir os resultados falso-negativos e falso-positivos. A partir da comparação de múltiplas sequências, é possível determinar regiões que possuem agrupamentos de genes diferentes entre espécies relacionadas (LANGILLE; HSIAO; BRINKMAN, 2010).

Se dois organismos possuem o mesmo ancestral comum, as possíveis diferenças entre os genomas são derivadas a partir do mesmo. Quanto mais estreitamente relacionados forem esses organismos, mais as suas sequências irão permanecer conservadas. Com isso, é possível determinar regiões divergentes entre os genomas (LANGILLE; HSIAO; BRINKMAN, 2008).

FIGURA 6. EXEMPLO DE ANÁLISE A PARTIR DE COMPARAÇÃO GENÔMICA



FONTE: Adaptado de CHE; HASAN; CHEN, 2014
NOTA: Três genomas selecionados como referência (G1, G2, G3) e suas respectivas regiões conservadas destacas em cores (verde região 1, azul região 2, amarelo região3, marrom região 4), para análise comparativa com o genoma de estudo G4. Em destaque em vermelho a região candidata a GI.

A análise, segundo os autores, consiste em três etapas principais, como exemplificado na figura acima (figura 6), são elas: 1) buscar genomas de referência ou estreitamente relacionados de acordo com organismo de interesse para estudo; 2) realizar alinhamentos múltiplos entre os genomas e 3) identificar regiões divergentes, indicando que este segmento pode ser proveniente de uma origem estrangeira, candidata a GIs. Ferramentas que utilizam essa abordagem, IslandPick, MobilomeFINDER, MOSAIC. (CHE; HSAN; CHEN, 2014). Figura 7 exemplifica uma comparação genômica entre organismos.

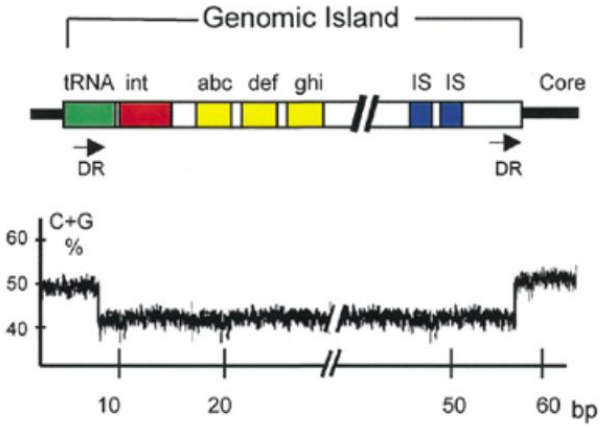
2.5.2 Análise de composição de sequência

Várias ferramentas têm se mostrado eficazes para identificar GIs a partir da composição de sequência dos organismos, entretanto os resultados podem conter algumas predições falso-negativos e falso-positivos. A grande diferença dessa abordagem em relação a genômica comparativa é que não requer genoma de referência para sua análise e comparação, que em muitos casos não se encontram presentes (LANGILLE; HSIAO; BRINKMAN, 2010).

Devido a vários fatores como, pressões mutacionais e forças de seleção agindo constantemente nos genomas bacterianos, essas condições podem levar a variações na composição de sequencias do organismo, sendo estas uma das características importantes neste tipo de abordagem (LEE et al., 2013). Levando em conta estes aspectos é possível identificar regiões atípicas que diferem do restante do genoma. Essas regiões apresentam composições de sequências com atributos distintos, como por exemplo, variações do conteúdo G+C e dinucleotídeos, frequência da utilização de códons, presença de genes relacionados a mobilidade, sequências de inserção (IS), repetições diretas, genes altamente expressos (CHE; HSAN; CHEN, 2014).

Com base nesses critérios foram desenvolvidas ferramentas para identificar essas possíveis regiões atípicas a partir da variação em sua composição de sequência. A maioria dos *softwares* designam pontuações ou valores limiares para cada gene ou região de interesse em seus resultados, podendo ser pré-definidas ou de forma dinâmica). De acordo com os valores estabelecidos, pontuações para mais ou para menos daquele limiar podem ser definidas como candidatas a GIs. Exemplo de ferramentas que utilizam essas abordagens, Alien Hunter, IslandPath-DIMOB, SIGI-HMM, Predict Bias (LANGILLE; HSIAO; BRINKMAN, 2010; LU; LEONG, 2016). Um exemplo simplificado dessas regiões e sua composição atípica presente no genoma pode ser visto na figura 7.

FIGURA 7. EXEMPLO DE UMA REGIÃO ATÍPICA E SUA COMPOSIÇÃO DISTINTA



FONTE: Adaptado de HACKER; CARNIEL, 2001.
NOTA: Representação de uma região candidata a GI com suas características distintas. Teor de citosina e guanina (C+G%) abaixo em relação ao restante do genoma, gene de tRNA seguido por repetições diretas, gene ligado a mobilidade e execisão integrase, genes representativos e sequencias de inserção.

2.6 FERRAMENTAS DE PREDIÇÃO DE ILHAS GENÔMICAS

2.6.1 Alien Hunter

Desenvolvido por pesquisadores do Instituto Sanger no Reino Unido, este software para predição de ilhas genômicas é baseado em Motivos de Ordem Variável Interpolados (IVOMs), que consiste em detectar regiões atípicas no genoma através da utilização de análises de composição de sequência como variação do conteúdo G+C, presença de dinucleotideos e frequência de códons. As predições podem ser optimizadas com a utilização de Modelos Ocultos de Markov de dois estados (HMM) para identificar o ponto de mudança entre regiões atípicas e não atípicas do genoma (VERNIKOS; PARKHILL, 2006).

Quando ocorre a identificação dessas regiões é obtido IVOM *score*, que equivale a quanto essa porção do genoma difere do restante. Sequências mais longas apresentam *scores* mais elevados e predições mais precisas, enquanto sequências menores e com poucas informações têm *score* mais baixo e com um resultado duvidoso (CHE; HASAN; CHEN, 2014). *Threshold* também é estabelecido, sendo ele um limiar com um *score* resultante da comparação com a média do genoma total relacionado a sua similaridade. Genes ou regiões genômicas com um *score* abaixo ou acima do limiar são possivelmente atípicos, genes subsequentes ou até mesmo essas regiões atípicas são unidos para obter candidatas a GIs (LU; LEONG, 2016).

Alien Hunter é capaz de realizar predições sem que seja necessária uma anotação pré-existente, ou informações que contenham as posições que cada gene se encontra no genoma. Sendo assim, ele pode ser utilizado em genomas recém sequenciados (CHE; HASAN; CHEN, 2014).

É um software livre de código aberto que pode ser redistribuído e modificado. O Sistema Operacional da ferramenta é somente Linux e não possui nenhum requerimento de *hardware* computacional descrito seja mínimo ou recomendado para rodá-lo. A versão equivalente ou superior da linguagem de programação Pearl v5.6.1 e Java SDK v1.4.2. se faz necessária, sendo a ferramenta executada por linha de comando.

Seu arquivo de entrada é em formato .FASTA. Este formato compreende as sequências de nucleotídeos do genoma, não exigindo pré-processamento dos dados.

Não há necessidade de baixar ou se conectar a qualquer banco de dados para realizar as análises.

Os arquivos de saída da ferramenta possuem extensões variadas, sendo que podem ser utilizados em conjunto com a ferramenta de visualização de genomas Artemis (CARVER et al., 2012) para melhor compreensão dos resultados. O primeiro em .TXT, possui informações de todas as regiões atípicas, com sua posição inicial e final no genoma de cada GIs candidata, juntamente com seu IVOM *score* e *Threshold*, além de informar a coloração das *Features* correspondente para visualização na ferramenta externa Artemis. A segunda extensão .PLOT é um arquivo para plotagem de gráfico correspondente às regiões atípicas no genoma que a ferramenta detectou, podendo ser inserida como arquivo de entrada no software Artemis.

2.6.2 GI Hunter

Esta ferramenta foi desenvolvida em *East Stroudsburg* pelo laboratório de Bioinformática da Universidade da Pensilvânia. É capaz de identificar ilhas genômicas tanto em genomas bacterianos quanto archaea. É baseado em análises de composição de sequência, genes de tRNA e genes altamente expressos, distância intergênica, informações sobre fagos e genes móveis (integrase e transposases), além da implementação da metodologia de Motivos de Ordem Variável Interpolados (IVOMs) que a ferramenta Alien Hunter utiliza para realizar suas análises (CHE; WANG; FAZEKAS, 2014).

Para prever as ilhas genômicas foi desenvolvido também um método de predição baseado em árvore de decisão com um conjunto de treinamento. Os atributos dos genes altamente expressos e as distâncias intergênicas não foram explorados em outras ferramentas, podendo ocorrer uma possível melhoria nos resultados das predições (CHE; HASAN; CHEN, 2014).

O conjunto de treinamento foi obtido a partir de 118 genomas do estudo de Langille e colaboradores (*Evaluation of genomic island predictors using a comparative genomics approach* – IslandPick), sendo que neles constavam 771 GIs supostamente positivas e 3,770 GIs negativas. As ilhas genômicas variaram muito de tamanho, não sendo um atributo forte para a construção do modelo de predição para GI Hunter. Para tentar contornar esse problema foi retirado todos os genes das GIs candidatas, tanto das positivas quanto negativas (CHE; WANG; FAZEKAS, 2014).

Para cada gene obtido foram analisados diversos aspectos. Entre eles a pontuação IVOM para os genes selecionados dentro da região de interesse, através da implementação da metodologia da ferramenta Alien Hunter. Para genes de tRNA, fagos, integrase e transposase, buscaram informações contidas nas anotações dos genomas, identificando a quantidade desses genes que estariam presente nas regiões candidatas a GIs. Foram considerados genes altamente expressos, aqueles que possuem função de tradução ou de transcrição, genes envolvidos no metabolismo de energia e de proteína ribossômica que estivessem dentro da região de interesse (CHE; WANG; FAZEKAS, 2014).

A densidade dos genes foi definida de acordo com a quantidade de genes presentes das GIs candidata e a distância intergênica foi medida através de cada par adjacente dos genes de interesse. Com essas informações foi montado o modelo de classificação fundamentada em árvore de decisão utilizando o software WEKA - *Waikato Environment for Knowledge Analysis* (FRANK; HALL; WITTEN, 2016). Este modelo é um agregador de *bootstrap*, também conhecido como método de ensacamento (*bagging*), cuja classificação é baseada nos votos de cada classificador possuindo o mesmo peso para cada um. O conjunto de treinamento para a construção das árvores foi todos os genes retirados das GIs positivas e negativas, com as suas respectivas características designadas. Na utilização dos algoritmos para a construção do modelo foi empregado o algoritmo J48 do WEKA, sendo este a implementação Java do C4.8, um dos mais conhecidos para árvores de decisão, por sua capacidade de trabalhar com atributos em falta e também incorporar processos para melhorar os problemas de *over-fitting* (CHE; WANG; FAZEKAS, 2014).

Possui seu código fonte aberto para ser modificado e redistribuído, desenvolvido em linguagem de programação C++ e somente é executável em Sistema Operacional Linux. As versões equivalentes ou superiores do Java SDK v1.5 e Pearl 5.8 se faz necessária, mas não possui nenhum requerimento de hardware computacional descrito que seja mínimo ou recomendado para rodá-lo.

Os arquivos de entrada devem ser em formatos .FNA, .PTT, .RNT. A extensão .FNA corresponde as sequências em nucleotídeos do genoma em formato FASTA, já a extensão .PTT designa todas as proteínas presentes no genoma contendo informações como posição inicial, final e tamanho da proteína. Além disso, determina sua localização no genoma, contendo o sentido da fase de leitura que esse produto

se encontra: *Forward*, da esquerda para direita, e o *Backward*, leitura da direita para esquerda.

As informações contidas na extensão .RNT, ao invés de ser atribuído dados sobre proteínas essa extensão descreve todos os tRNAs contidos no genoma. A ferramenta não necessita de um pré-processamento dos arquivos de entrada para realizar as análises e também não a conexão ou a necessidade de baixar qualquer banco de dados para se obter os resultados.

Os arquivos de saída da ferramenta possuem somente duas extensões. O primeiro em .TXT, que contém apenas informações das posições inicias e finais das GIs candidatas, sendo ausente de qualquer outra informação como por exemplo IVOM *score* ou qualquer outro atributo estabelecido para classificar e modelar o método de predição. A segunda extensão em .ARFF possibilita a utilização de uma ferramenta externa, desenvolvida pelo mesmo laboratório, para visualização dos resultados provenientes da ferramenta GI Hunter, chamado de GIV (*Genomic Island Visualization*) (CHE; WANG, 2013). Com esta ferramenta é possível identificar somente suas posições das GIs candidatas sobre um genoma circular do organismo de estudo, sendo ausente a descrição dos produtos que se encontram dentro das Ilhas de interesse.

Outra ferramenta externa possível de ser utilizada é Artemis. Com as informações contidas na extensão TXT é possível localizar as posições das ilhas dentro do genoma e realizar estudos mais aprofundados sobre as funções biológicas dos genes que se fazem presente naquela região.

2.6.3 Genomic island prediction software - GIPSy

GIPSy é uma atualização da ferramenta PIPS (*Pathogenicity Island Prediction Software*), desenvolvido para identificar GIs patogênicas em genomas bacterianos. Após o melhoramento do software, o GIPSy é capaz de identificar tanto ilhas de patogenicidade (PAI) quanto outras regiões candidatas, bem como classificá-las de acordo com os genes presentes nas Ilhas em relação com suas funções biológicas (Mls, Rls, Sls). Para realizar as análises, GIPSy se baseia no desvio do conteúdo G+C e utilização de códons do genoma, genes de tRNA e de mobilidade como transposase, fatores de virulência, metabolismo, simbiose, resistência antibióticos que estejam

ausentes em outros organismos do mesmo género ou espécies. Para isso a ferramenta passa por oito passos de execução (SOARES et al., 2016).

No primeiro passo ocorre o processamento dos dados do genoma através do arquivo de entrada sendo necessário o mesmo ser em extensão .EMBL ou .GENBANK (.GBK, .GB). A partir de todas as sequências de codificação (CDS) do genoma provenientes do arquivo de entrada são gerados outros três arquivos com extensão .FNA, .FAA, .FFN, a fim de serem utilizados em processos futuros. Se o usuário fizer uso do arquivo .GBK para sua entrada será gerado também uma extensão .EMBL para ser utilizado obrigatoriamente no passo três pelo Colombo/SIGIHMM (WAACK et al., 2006).

No segundo passo, é analisado o desvio do conteúdo G+C do genoma de acordo com abordagens utilizadas no desenvolvimento da ferramenta anterior PIPs. A partir do arquivo .FNA gerado através do processamento dos dados é calculado o conteúdo G+C sobre o tamanho do genoma. Em seguida cada CDS proveniente do arquivo .FFN é analisado para determinar o conteúdo G+C das CDS pelo tamanho de todas as CDS. Com essas informações é determinado então o valor médio sendo considerado a porcentagem G+C do genoma, e o valor do desvio padrão (SD) obtido através do número total de CDS presente no genoma (SOARES et al., 2012).

No terceiro passo, é implementado o software Colombo/SIGI-HMM para analisar o desvio do uso de códons utilizando o arquivo com extensão .EMBL gerado no processamento dos dados. A ferramenta faz uso das Cadeias Ocultas de Markov (HMM) para medir o desvio da utilização de códons através de uma busca a partir da composição de sequência do genoma (WAACK et al., 2006).

No quarto passo, acontece a busca por genes transposase que é realizado com a implementação do software HMMER3, sendo sua análise efetuada através do arquivo com extensão .FAA gerado no começo do processamento de dados. HMMER3 é capaz de efetivar buscas de sequências homólogas em bancos de dados e realizar alinhamentos de sequências com uma velocidade tão rápida quanto as buscas por BLAST. Esta velocidade é devido ao algoritmo descrito como Segmento Múltiplo de Viterbi (MVS), sendo uma heurística de aceleração de Perfil-HMM calculando o menor caminho provável para composição de uma sequência de DNA (EDDY, 2011). O banco de dados para busca do gene transposase é realizado através PFAM - *Protein Families Database*. Inúmeras famílias de proteínas se encontram

depositadas neste banco e cada uma delas é representada por múltiplos alinhamentos de sequência e Modelos Ocultos de Markov (FINN et al., 2013).

No quinto passo, a classe de cada GI candidata é determinada através de cinco banco de dados, sendo cada um deles específico para cada fator. O arquivo com extensão .FAA gerado no processamento de dados é utilizado para pesquisa de similaridade de proteínas em cada um dos bancos a partir do BLASTP (SOARES et al., 2015). A busca por genes relacionados a virulência é realizado no banco de dados MvirDB - *Microbial Database of Protein Toxins, Virulence Factors*. Este banco contém sequências de DNA e proteínas provenientes de outros bancos, entre eles, Tox-Prot, SCORPION, PRINTS, VFDB, TVFac, Islander, ARGO, CONUS, VIDA (Zhou et al., 2007). A busca pelos fatores de resistência a antibióticos ocorre em dois bancos de dados ARDB - *Antibiotic Resistance genes Database* (LIU; POP, 2009) e CARD - The *Comprehensive Antibiotic ResistanceDatabase* (MCARTHUR et al., 2013). O banco de dados para os fatores relacionados ao metabolismo foi criado a partir de todas as proteínas ortólogas resultantes de todos os genes presentes nos clusters da categoria de metabolismo do banco de dados COG - *Clusters of Orthologous Groups of proteins* (SOARES et al., 2016). Por fim, na busca de fatores relacionados a simbiose, o banco de dados NodMutDB - *Nodulation Mutant Database* é utilizado. Neste banco constam genes coletados através de pesquisa em banco de dados públicos e revisão de literatura sendo que a maioria dos dados é voltado principalmente para genes de fixação de nitrogênio (MAO et al., 2005).

No sexto passo, BLASTs recíprocos são executados entre todas as CDS do genoma de estudo e também no de referência, afim de identificar possíveis regiões que contenham similaridade e genes putativos, que se encontram presentes no genoma de estudo e ausentes no genoma de referência (SOARES et al., 2016).

No sétimo passo, os genes de tRNA são localizados a partir do arquivo de extensão .FNA gerado no processamento de dados inicial, utilizando novamente a implementação do software HMMER3 para busca no banco de dados tRNAdb – *Transfer RNA Database* contendo compilações de sequencias de tRNA e genes de tRNA (JUHLING et al., 2009).

No último processo, após todas as análises anteriores serem realizadas, esta etapa faz a junção de todas as informações adquiridas em um arquivo delimitado por tabulações. De acordo com os dados provenientes do BLAST, as CDS que possuem similaridade são agrupadas em clusters de genes que são compartilhados por ambos

os genomas, de estudo e referência. Posteriormente, as regiões que apresentam tamanho maior que 6 kb e que foram localizadas apenas no genoma de estudo, são escolhidas como GIs putativas, devendo conter as combinações de características analisadas maiores do que aquelas apresentadas em toda a sequência do genoma (SOARES et al., 2016).

Esta ferramenta foi desenvolvida em JAVA podendo ser executada em Sistema Operacional Windows e Linux, possuindo interface gráfica em ambos os sistemas. Requer versão do Java *Virutal Machine* 1.7.0_51-b13 ou superior. Em sistemas Linux é recomendado fazer uso do *Openjdk* substituindo a versão *Oracle*, devido algumas exceções que podem ocorrer durante as análises. Não possui nenhum requerimento de hardware computacional descrito seja mínimo ou recomendado para rodá-lo. Os arquivos de entrada para as análises são genomas completos um para estudo e outro para referência com extensão .EMBL ou .GBK (.GENBANK, .GB).

Com os arquivos de entrada, a ferramenta realiza um pré-processamento dos dados para gerar outros três arquivos necessários para utilizar no decorrer das análises, sendo eles: .FNA (FASTA nucleotídeos), .FAA (FASTA aminoácidos) e .FFN (FASTA de nucleotídeos codificados por região). Uma extensão .EMBL também é gerada se o usuário fazer uso do formato .GBK como entrada, devido a ferramenta implementada chamada Colombo/SIGIHMM aceitar somente este tipo de arquivo para realizar as análises. Todas as dependências necessárias, como os softwares implementados BLAST, COLOMBO/SIGIHMM, HMMER3 e os bancos de dados de virulência, metabolismo, resistência, simbiose, inclusive os de procura para genes de transposase e tRNA, são baixados e compilados automaticamente com o instalador da GIPSy, sem que ocorra a necessidade do usuário configurar cada um deles.

Os dados de saída da ferramenta correspondem a cada passo das análises do primeiro até ao oitavo processo. Do segundo passo em diante, todos os arquivos de saída possuem extensão .TXT. No resultado do primeiro passo é possível obter diversas extensões de arquivos provenientes dos dados de entrada, tanto do genoma de estudo quanto o de referência, sendo eles .FAA, .FNA, .FFN, .EMBL, .GBK.

No segundo passo é possível visualizar os resultados da pesquisa do desvio do conteúdo G+C, mostrando o conteúdo total e porcentagem de G+C, o tamanho do genoma, os valores de referência usados pela ferramenta de análise juntamente com os limites inferiores, superiores e o desvio em cada CDS correspondente no genoma de estudo e de referência em extensão.

No terceiro passo, o resultado da análise de preferência de códons é demonstrada com os parâmetros utilizados pela ferramenta de busca COLOMBO/SIGIHMM, e pode-se observar cada CDS presente no genoma de estudo e referência se ocorreu desvio ou não.

No quarto passo, a procura pela presença de transposase ocorre somente no genoma de estudo. Os resultados são apresentados com os parâmetros utilizados pela ferramenta HMMER3 e a identificação de cada CDS correspondente com a procura do gene de interesse.

No quinto passo, são apresentados os resultados da pesquisa por BLAST nos bancos de dados relacionados (Metabolismo, Resistência, Simbiose, Virulência), para classificar as GIs candidatas de acordo com os genes presentes em cada ilha, relacionados a suas funções biológicas. É possível dessa forma visualizar cada CDS similar aos dados encontrados pelo BLAST.

No sexto passo, todas as CDS resultantes da procura pelos BLASTs recíprocos entre os genomas são apresentandos com a sua porcentagem de similaridade.

No sétimo passo, os resultados derivados pela busca de tRNA são apenas do genoma analisados pela ferramenta HMMER3, sendo apresentado de acordo com suas posições inicias e finais no genoma, além de apresentar a direção da fase de leitura. Ainda traz informações sobre tRNA truncados e os parâmetros utilizados pela ferramenta de busca.

No último passo, os resultados apresentam as ilhas já determinadas, contendo suas respectivas classes e várias informações sobre as características de cada ilha, entre elas: a composição dos genes de cada ilha, sua locus tag inicial e final, a posição das ilhas no genoma com seu início e fim, o quanto aquela predição é significativa, sendo forte, normal ou fraca. Além disso, mostra várias características de uma das ilhas como desvio de G+C, uso de códons, proteínas hipotéticas e porcentagens da presença de genes relacionados a cada classe, de acordo com seus fatores de virulência, metabolismo, simbiose ou resistência.

2.6.4 IslandViewer3

Este software web foi desenvolvido na *Simon Fraser University*, pelo laboratório *Brinkman Lab* no Canada. Também é um banco de dados de Ilhas genômicas

contendo organismos bacterianos e archaea. Para realizar as predições, o IslandViewer3 faz uso de três metodologias integradas, são elas: IslandPick, que utiliza comparação genômica, SIGI-HMM para pesquisa de composição de sequência e IslandPath-DIMOB, que busca sequências atípicas e genes relacionados a mobilidade. Anotações externas provenientes de outros bancos de dados relacionados com fatores de virulência, patogenicidade, resistência a antibióticos e genes homólogos a este fator também estão disponíveis (DHILLON et al., 2015).

Uma das metodologias integradas é a IslandPick, que faz uso de comparações genômicas, sendo necessário para suas análises genomas filogeneticamente relacionados. A partir de um único genoma de estudo, esta metodologia é capaz de separar tanto dados positivos a serem candidatas a GIs, quanto negativos. Utilizando CVTree (XU; HAO, 2009) que constrói árvores filogenéticas baseadas em genomas completos sem alinhamento de sequências usando uma abordagem de Vector de Composição (CV), determina-se uma distância pré-calculada a partir do genoma de estudo, atribuindo pontos de cortes de distância para poder selecionar genomas de referência de espécies ou estirpes estreitamente relacionadas. Se os genomas de referência forem suficientes para serem utilizados na comparação, então os genomas tanto de estudo quanto de referência são analisados no MAUVE (DARLING et al., 2004). O MAUVE é um software é capaz de realizar alinhamentos múltiplos de genomas e posteriormente, todas as regiões conservadas entre os dois genomas são extraídas como conjunto de dados negativos. A construção dos conjuntos de dados positivos é obtida através do emparelhamento do genoma de estudo com os genomas de referência, sendo que todas as regiões não alinhadas passam por verificação no BLAST para se ter certeza que esta seja exclusiva do genoma de estudo, atribuída então como uma positiva candidata a GI (LANGILLE; HSIAO; BRINKMAN, 2008).

SIGI-HMM é um método de predição de ilhas genômicas em procariotos baseado em composição de sequência a partir de Modelos Ocultos de Markov (HMM). Este método de predição mede a frequência de códons utilizado em cada gene do genoma. Para cada um desses genes a utilização da frequência de códons é comparado com uma tabela contendo genes altamente expressos e doadores microbianos. A tabela com as informações das frequências de um organismo pode ser proveniente do seu próprio genoma se estiver disponível ou é medida através da tabela do banco de dados CUTG - *Codon Usage Database*. Vários testes são executados múltiplas vezes para identificar genes atípicos e para prever supostos

genes altamente expressos. Para esses genes, a busca pela utilização de códons é medida com a sua própria tabela (organismo hospedeiro), e a mesma sequência do gene é analisada, mas agora com a tabela do possível organismo doador, a fim de identificar as suas frequências (WAACK et al., 2006).

O IslandPath-DIMOB identifica sequências que sejam atípicas no genoma e genes de mobilidade para determinar seus resultados. Para análises da porcentagem de conteúdo G+C é utilizado uma única sequência de fase de leitura aberta (ORF) em conjunto com um agrupamento de ORFs, permitindo investigar a mudança da variação de seu conteúdo, tanto gene por gene quanto gene por clusters entre as supostas sequências de codificações. Outra busca por sequência anormal também considera é o viés de dinucleotideo, sendo uma assinatura de DNA independente que não se enquadra na mesma forma de avaliar a variação do conteúdo G+C.

Para determinar esse viés de dinucleotideo foi implementado uma adaptação das fórmulas de análise de Samuel Karlin (2001), que resumidamente divide todo o genoma em cluster de seis ORFs consecutivas para efetuar os cálculos tanto dos cluster quanto das regiões. Para genes de mobilidade, tais como integrases e transposases, são derivadas das anotações dos genomas, e a presença de tRNAs é proveniente da implementação do *software* tRNAscan-SE para sua identificação (HSIAO et al., 2003).

Anotações externas de genes relacionadas a vários fatores são derivados de outros bancos e se encontram implementados nos resultados finais. Para fatores de virulência, a informação sobre os genes é obtida através de três banco de dados. O primeiro VFDB – *Virulence Factor Database* (CHEN et al., 2012), para genes voltadas a fatores de virulência; O segundo contém genes curados manualmente denominado de VICTOR'S – *Virulence Factors*; O terceiro possui tanto fatores de virulência quanto de resistência a antibióticos chamado de PATRIC - *Pathosystems Resource Integration Center* (WATTAM et al., 2014). Para os genes relacionados com resistência a antibióticos e genes homólogos a este fator, se faz uso do banco CARD - *Comprehensive Antibiotic Resistance Database* (MCARTHUR et al., 2013). Por fim, para fatores de patogenicidade é utilizado um suplemento web contendo inúmeras anotações de genes relacionados (HO SUI et al., 2009).

Esta ferramenta possui uma interface integrada para identificação computacional e visualização de ilhas genômicas via Web, sendo também um banco de dados contendo GIs de organismos bacterianos e archaea.

Todas as páginas web foram escritas em PHP com *scripts* em Pearl para rodar processos dinâmicos, não possuindo restrição de navegadores de internet.

Para armazenar os dados pré-computados das GIs, o *IslandViewer3* faz uso de servidores *MySQL*. Esses dados são provenientes de todos os genomas completos depositados no NCBI - *National Center for Biotechnology Information* utilizando o MicrobeDB, o qual é atualizado conforme novos dados sejam depositados. Todos os métodos de predição são rodados em paralelo para cada novo genoma depositado ou atualizado, para que os processos sejam executados rapidamente.

Os arquivos de entrada para a realização das predições devem ser inicialmente enviados para o banco, através do upload do genoma de interesse para estudo. Os formatos aceitos são .GBK e.EMBL. O usuário pode determinar o nome do genoma de teste para identificá-lo melhor e atribuir um e-mail para receber os resultados após processamento, porém é opcional. Com a ausência de qualquer endereço de e-mail para recebimento dos dados, é possível acompanhar os processos através do endereço da URL (Localizador Uniforme de Recursos) aberta logo após a aceitação do genoma para análise. Nesta página o usuário pode visualizar, a partir de um fluxograma, em que momento se encontra o seu genoma nas ferramentas de predição, podendo estar pendente, processando, completo ou com erro. Todos os resultados são armazenados no próprio banco do IslandViewer3 por um período de um mês, contando a partir do dia da submissão do genoma. *Drafts* de genomas também são aceitos para análise, é recomendado diminuir o máximo número de *contigs* possíveis para reduzir falsas predições e perdas de GIs. A ordenação dos *contigs* será realizada através do Mauve Contig Mover (RISSMAN et al., 2009).

Os arquivos de saída da ferramenta são variados podendo ser visualizados na própria página web e algumas informações podem ser baixadas. Um dos resultados é uma plotagem interativa do genoma de estudo, tanto circular como horizontal, através do *software* implementado IslandPlot, desenvolvido em D3 *Javascript Library*. Os usuários podem escolher qual metodologia de predição e anotações externas devem ser exibidas na plotagem. Ao clicar nas GIs candidatas presentes na imagem, é atualizado o genoma horizontal na região de interesse, mostrando todos os genes presentes naquela ilha com a descrição do método de predição, fatores a ele relacionados se houver (virulência, resistência, patogenicidade), nome do gene, número de acesso no NCBI com link direto e nome do produto. Logo abaixo em uma tabela se encontra informações como posição inicial e final da ilha no genoma, o seu

tamanho, o método de predição utilizado divididos por cores (vermelho corresponde aos métodos integrados, em verde IslandPick, laranja SIGI-HMM, azul IslandPath-DIMOB) e um link direto para a plotagem interativa do genoma circular mostrando as informações dos genes descritas anteriormente.

É possível baixar todas as informações contidas nas ilhas separando por métodos de predição ou integrados em formato tabulado, contendo a descrição da posição inicial e final da ilha, o seu tamanho, o método de predição, nome do gene com sua *locus tag*, posição e sentido na fase de leitura (*Forward/Bacward*). As anotações provenientes de outros bancos também estão presentes nos mesmos formatos, trazendo dados sobre cada CDS relacionado com alguns dos fatores de interesse, contendo o número de acesso no NCBI, nome e link do banco com o produto encontrado descrevendo características daquele gene, função e referencia bibliográfica. As sequências das ilhas podem ser adquiridas em formatos .GBK e também em .FASTA podendo ser separados de acordo com os métodos de predição ou em conjunto.

O banco ainda disponibiliza a opção de baixar os dados de cada ilha separadamente. A imagem do genoma circular e horizontal também pode ser exportada em formato .PNG e .SVG. Além disso, o Islandviewer3 possui uma Toolbox na página de resultados que pode ser utilizada pelo usuário para procurar por genes específicos, visualizar dois genomas para estudo e comparação, e opção para escolher o genoma de referência utilizado na análise pelo método IslandPick. Isto mostra automaticamente organismos estreitamente relacionados, podendo ser a critério do pesquisador.

2.6.5 Zisland Explorer

Desenvolvido em Tianjin *University BioInformatics* Centre na China, o Zisland Explorer utiliza estratégias diferentes para predição de ilhas genômicas. É uma ferramenta de anotação não supervisionada e dependente de algoritmo para segmentação automatizado. O Zisland Explorer implementa o *software* GC+Profile para dividir toda a sequência do genoma em vários fragmentos para posterior análise. Essa abordagem combina homogeneidade de sequencias dentro de cada ilha e heterogeneidade das composições de sequência (WEI et al., 2016).

Com as possíveis variações na composição de bases ao longo de todas as sequencias codificantes e não codificantes, a estrutura genômica dos organismos podem ser afetadas. GC-Profile amplia o conceito de organização do genoma elaborando um novo método para segmentação baseado na divergência das sequencias. Com as posições dessas sequencias divergentes é possível gerar uma representação gráfica para estipular a variação do conteúdo G+C permitindo assim estabelecer relações biológicas funcionais dos genes presentes naquela região (GAO; ZHANG, 2006).

O Zisland Explorer adota uma estratégia de três fases de análise para identificar as GIs candidatas. A primeira fase, consiste no processo de segmentação de toda a sequência do genoma de acordo com sua homogeneidade do conteúdo G+C. Com os fragmentos de DNA identificados, a próxima fase é identificar os dados pertencentes ao *core* genoma antes de realizar a predição das GIs. Para isso, a segunda fase é subdividida. No primeiro momento, ocorre a criação de cluster contendo os fragmentos do *core* genoma utilizando a heterogeneidade do conteúdo G+C entre eles e todo o genoma. Já no segundo momento, outro cluster é gerado, mas sendo resultante da análise da homogeneidade do conteúdo G+C com o do *core* genoma que se encontra dentro de cada segmento. Cada cluster contendo os fragmentos do *core* genoma possui menos heterogeneidade e homogeneidade de conteúdo G+C. Obtidas essas informações as GIs são então identificas possuindo um ponto de corte em relação ao seu tamanho, mantendo somente ilhas com um comprimento entre 2 e 400 kb no resultado das predições (WEI et al., 2016).

Este software foi desenvolvido para múltiplas plataformas com interface gráfica, podendo ser executado em Windows, Linux e Mac possuindo também uma aplicação *Web*. Nenhum requerimento de hardware computacional se encontra descrito que seja mínimo ou recomendado para rodá-lo *in house* e os navegadores de internet para análises via *Web*, não possuem restrições. A ferramenta necessita de alguns módulos essenciais de linguagem de programação para instalação, são eles: Python 2.0 ou superior, Biopython e Matplotlib. É ausente de qualquer pré-processamento dos dados, assim como a necessidade de baixar ou se conectar a qualquer banco de dados para realizar as predições.

Os arquivos de entrada devem ser em extensão .FNA em formato FASTA, com as sequencias do genoma em nucleotídeo, e um arquivo com todas as informações

relacionados as proteínas presentes no organismo de estudo derivada da extensão .PTT.

Os arquivos de saída se apresentam de duas formas tanto para as ferramentas *in house* quanto aplicação web. O primeiro, em extensão .TXT, possui informações como posição inicial e final da GI no genoma, seu tamanho correspondente, o número de genes presentes dentro da GI e um *score* determinando o quanto aquela região possui de heterogeneidade e homogeneidade de conteúdo G+C. E o segundo arquivo é uma plotagem em gráfico de linha identificando a variação de heterogeneidade e homogeneidade de conteúdo G+C de cada ilha predita no decorrer do genoma.

2.6.6 Predict Bias

Este software foi desenvolvido no laboratório de bioinformática da Devi Ahila University, Indore na Índia capaz de identificar ilhas genômicas e de patogenicidade em organismos procariotos a partir da avaliação de composição de sequência, presença de elementos de inserção, genes relacionados a fatores de virulência e ausência dos mesmos em espécies não patogênicas. Para identificar os genes com essas características, foi criado um banco de dados interno, o VFPD (*A profile database of virulence factors*), com o objetivo de buscar a presença desses genes no genoma através da execução do RPS-BLAST (*Reversed Position Specific - Basic Local Alignment Search Tool*) nas regiões de interesse, cruzando as informações encontradas com o banco (PUNDHIR; VIJAYVARGIYA; KUMA, 2008).

Para as predições de genes tRNA e mobilidade como integrases e transposases, o Predict Bias faz uso das anotações do próprio genoma para determiná-los. Nas análises de composição de sequência como conteúdo G+C, dinucleotideos e códons é realizado através de cluster de seis ORFs. Esses cluster são formados a partir do modelo de janela deslizante (*sliding window shifting*) baseado na leitura de uma ORF por vez para estimar cerca de 1.500 codons ou 4,5 kb, que irão corresponder a cerca de 6 a 8 ORFs para montar o cluster. Com essas informações é possível estimar o desvio do conteúdo G+C ao longo do genoma. Para os cálculos relacionados a composição de dinucleotideo e códons, os autores implementaram os métodos e fórmulas desenvolvidos por Samuel Karlin (2001). As regiões candidatas a GIs são denotadas então a partir de cluster consecutivos contendo seis ORFs com valor acima do limiar sobre o desvio do conteúdo G+C e composição dos dinucleotideo

e códons (PUNDHIR; VIJAYVARGIYA; KUMA, 2008). O valor limiar foi obtido através da análise de setenta e três ilhas genômicas provenientes do banco de dados Islander a partir de cinquenta e dois genomas bacterianos depositados (MANTRI; WILLIAMS, 2004).

A classificação das ilhas de patogenicidade é feita pela comparação da presença de genes relacionados a fatores de virulência nas regiões candidatas a GI, a partir da pesquisa por similaridade entre esses genes e o banco de dados VFPD. Os dados de todas as famílias de proteínas relacionados com virulência são originários do banco de dados PFAM e PRINT pesquisando por palavras chaves como '*Virulence', 'Adhesin', 'Siderophore', 'Invasin', 'Endotoxin',* e 'Exotoxin'. Se apenas um gene se encontrar presente dentro do banco, essa GI passa a ser classificada como PAI. Regiões que não atingiram o valor limiar das composições de sequência, mas possuem genes similares depositados no banco, irão ser classificadas como PAI- sem composição de sequência (*unbiased composition*) (LANGILLE; HSIAO; BRINKMAN, 2010).

Esta ferramenta é uma aplicação web desenvolvida por Devi Ahilya, no laboratório de Biotecnologia da Universidade Indore, Índia. Foi implementada utilizando ASP.Net e linguagens de programação como Pearl e C++, rodando em *IIS web server*, utilizando o *SQL server database*. Não possui restrições de navegadores de internet para utilização. Esta aplicação trabalha ainda, de certa forma, como um banco de dados, armazenando os resultados dos genomas que já foram submetidos para análise, de modo que qualquer novo genoma enviado para estudo que não se encontre já depositado irá ser implementado no repositório.

O arquivo de entrada deve ser em extensão .GBK, contendo toda a sequência do genoma e suas devidas anotações. O usuário pode optar por alterar os valores de limiar do conteúdo G+C e composição de dinucleótideo e códons antes de submeter os dados para análise.

Os arquivos de saída da ferramenta na aplicação web estão dispostas pela identificação a partir da composição de sequência (*biased composition*), seguidas das ilhas identificadas, e também dados das predições que não atingiram o valor limiar necessário, porém contendo genes de virulência, sendo classificadas como (*unbiased composition*). Os resultados demonstram o início e final da ilha no genoma pela posição da *locus tag* correspondente, com a opção de estender as informações de cada ilha, é possível ser visualizado todos os valores da alteração de composição

identificados para cada CDS e seu nome, juntamente com os genes identificados pelo banco VFPD. Nesta fermenta consta ainda a possibilidade de realizar comparação de genomas, além de possuir uma árvore filogenética de espécies para pesquisa.

Para os arquivos de saída que podem ser baixados da aplicação, consta um arquivo em extensão .TXT e .HTML com todas as informações das ilhas descritas anteriormente e plotagens de gráficos em barras de todos os clusters identificados com suas respectivas alterações nas composições, em extensão .HTML e .PNG.

2.7 BANCO DE DADOS ILHAS GENÔMICAS

Há uma grande variedade de predições depositadas em bancos de dados que podem ser acessadas diretamente via web. Esses resultados podem ser comparados com os dados que o usuário obteve, utilizando as ferramentas de predição ou metodologias abordadas, para inferir qualidade dos dados obtidos ou depositados nos bancos. É versátil principalmente para aqueles usuários que não estão familiarizados com programas via de comando e linguagem de programação.

2.7.1 Database of Genomic Island – D.G.I

Este banco foi criado a partir dos resultados provenientes da ferramenta GI Hunter (CHE; WANG; FAZEKAS, 2014), desenvolvida pelos mesmos criadores do banco. Contém informações das ilhas genômicas e dados das predições, como posição inicial e final da ilha no genoma, assim como seus produtos presentes, IVOM (motivos de ordem variável e interpolados) *score* , densidade dos genes, distancia intergênica, genes altamente expressos, e presenças tRNA e elementos moveis, tais como integrase, transposase. Possui mais de dois mil genomas bacterianos depositados. Os dados de cada organismo são derivados do NCBI (Centro Nacional de Informações Biotecnológias). Para cada genoma depositado é gerado uma plotagem circular do genoma a partir do *software* GIV (*Genomic Island Visualization*) (CHE; WANG, 2014) representando cada ilha.

2.7.2 Islander

Possui cerca de quatro mil ilhas depositadas de genomas bacterianos, archaea e plasmídeos. Contém informações como o número de acesso do organismo no NCBI, taxonomia do organismo; tamanho da ilha, sítio de integração (presença de tRNA), conteúdo G+C, posição inicial e final da ilha no genoma e orientação, sequência em nucleotídeos da ilha, e descrição dos produtos presentes na região de interesse (HUDSON; LAU; WILLIAMS, 2015).

2.7.3 IslandViewer3

Dispõe de resultados pré-computados de genomas bacterianos e archea pelos métodos IslandPick (LANGILLE; HSIAO; BRINKMAN, 2008), IslandPath-DIMOB (HSIAO et al., 2003), e SIGI-HMM (WAACK et al., 2006) e também o usuário pode submeter o seu próprio genoma de estudo para análise. Contém um gráfico interativo do genoma circular, disponibilizando várias informações das ilhas dispostas na plotagem, como descrição dos genes presentes, número de acesso no NCBI, posição inicial e final da ilha. Possui também anotações externas de outros bancos, como VFDB – *Virulence Factor Database* (CHEN et al., 2012), VICTOR'S – *Virulence Factors*, PATRIC - *Pathosystems Resource Integration Center* (WATTAM et al., 2014), CARD - *Comprehensive Antibiotic Resistance Database* (MCARTHUR et al., 2013), para relacionar genes encontrados nas ilhas com fatores de virulência, resistência a antibióticos e patogenicidade. É possível baixar os dados das predições assim como a plotagem do genoma circular (DHILLON et al., 2015).

2.7.4 Pathogenicity Island Database - PAIDB

Este banco de dados é voltado para ilhas de patogenicidade (PAIs) e ilhas de resistência a antibióticos (REIs), sendo possível enviar genomas para análise. Contém cerca de 200 PAIs identificadas e 80 REIs a partir de dois mil genomas bacterianos originários do NCBI. Apresenta uma plotagem circular do genoma e vários dados sobre as ilhas, como o número de acesso no NCBI, conteúdo G+C, posição inicial e final no genoma e seu nome na literatura se disponível, tamanho da ilha, sitio de

inserção (tRNA correspondente), número de ORFs e sua descrição e regiões homologas em outros organismos (YOON; PARK; KIM, 2015).

2.7.5 Pré-GI

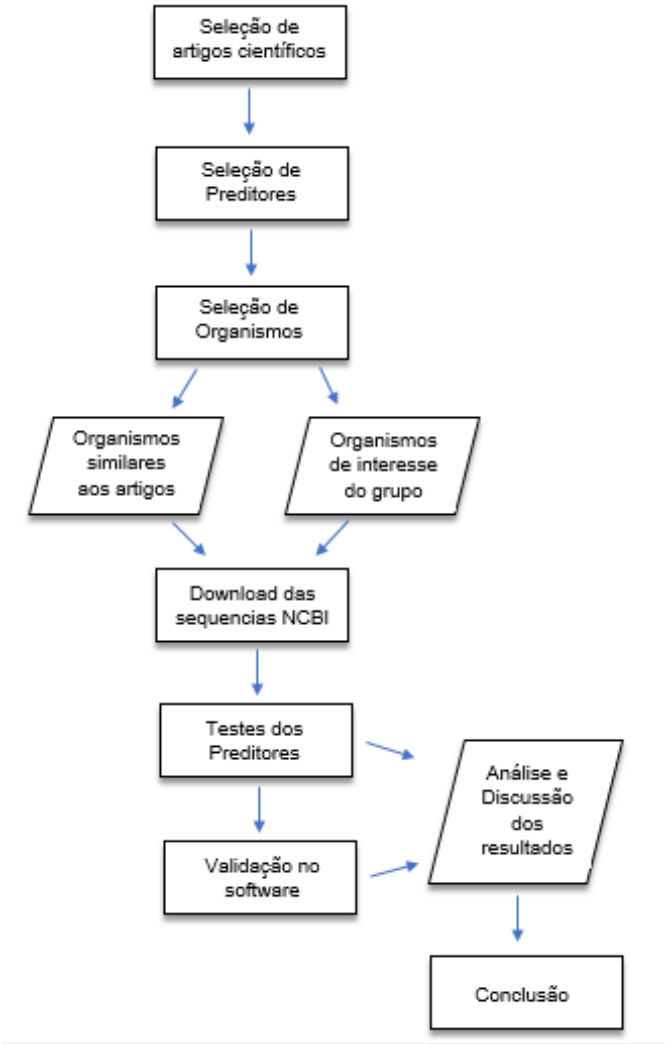
Compreende cerca de 26,000 ilhas identificadas em 2,000 genomas de bactérias, archea e plasmídeos, possibilitando ao usuário enviar seu genoma de estudo para análise. Na apresentação de seus resultados é possível visualizar uma plotagem do genoma circular interativo, contendo informações como, G+C, início e final da ilha no genoma, tamanho, descrição dos genes presentes, link para outros bancos (PAIDB, IslandViewer3), sequência da ilha em formato *genbank*. Este banco também provém de informações sobre a presença de ilhas genômicas em eucariotos, contendo algumas predições (PIERNEEF et al., 2015).

3 MATERIAIS E MÉTODOS

Este trabalho é de caráter qualitativo e quantitativo. Buscamos avaliar ferramentas de predições de ilhas genômicas utilizando conjunto de organismos, considerando seus diferentes resultados e buscando a melhor estratégia de predição.

Pesquisas de revisão bibliográficas são fundamentais para a comunidade científica, pois estas favorecem o esclarecimento de lacunas no conhecimento resultantes de diversas pesquisas já realizadas por outros autores (SAMPAIO, MANCINI, 2007).

FIGURA 8. PROCESSOS DE ELABORAÇÃO DA PESQUISA



FONTE: O autor (2017).
NOTA: Protocolo da pesquisa incluindo os itens de seleção, tratamento e validação dos dados, análise e interpretação dos resultados, seguido do processo final de conclusão.

A Figura 8 apresenta o *workflow* dos principais passos do processo desta pesquisa.

3.1 CRITÉRIOS PARA ESCOLHA DOS PREDITORES E SELEÇÃO DE ARTIGOS CIENTÍFICOS

Para a escolha das ferramentas de predição, os seguintes critérios foram estabelecidos: 1) é necessário que o *software* seja livre; 2) disponível para uso local ou *web*; 3) devendo ser independente/*stand alone* e; 4) publicado a menos de três anos ou com mais de vinte citações nos últimos três anos.

A tabela 1 descreve as ferramentas escolhidas, o número de citações dos artigos nos três últimos anos a partir do seu ano de publicação encontrados nas bases de dados *Google Scholar*, *Web of Science* e *Scopus*, na data de 07/03/2016.

TABELA 1 - NÚMERO TOTAL DE CITAÇÔES DAS FERRAMENTAS NOS ÚLTIMOS TRÊS ANOS

Preditores	Google Scholar	Web of Science	Scopus	Total	Ano de publicação
Alien Hunter	81	52	56	189	2006
GI Hunter	0	0	0	0	2014
GIPSy	0	0	0	0	2016
IslandViewer3*	216	147	122	485	2009,2013, 2015
Predict Bias**	4	2	2	8	2008
Zisland Explorer	6	0	0	6	2016

FONTE: O autor (2017).
NOTA: IslandViewer3* possui três artigos de publicação sobre atualização do banco e metodologias sendo indicados na tabela o número total de cada um. Predict Bias** foi incluído, mesmo não atendendo os critérios relativos ao número de citações e ano de publicação, devido a ferramenta GIPSy ter utilizado para comparação de resultados em sua publicação.

As ferramentas que não conseguiram atingir os critérios, seguidas de seu ano de publicação e número total de citações nos últimos três anos são: Centroid (2007) com 19 citações; GI Detector (2010) com 11 citações; INDeGenIUS (2010) com 10 citações; EGID (2011) com 12 citações; IGIPT (2011) com 4 citações; GIST (2012) com 17 citações.

3.2 CRITÉRIOS PARA ESCOLHA DOS ORGANISMOS TESTES

Para a escolha dos organismos, identificamos todos os conjuntos de testes utilizados pelos autores das ferramentas em suas publicações. Optamos por escolher organismos similares que foram testados em mais de uma ferramenta. Os organismos que possuem interesse para pesquisas junto ao laboratório de Bioinformática da UFPR (*Aeromonas hydrophila e Streptococcus pneumonae*), também foram selecionados.

Na tabela abaixo (tabela 2), estão retratados os organismos escolhidos para o presente trabalho, juntamente com uma breve descrição. A descrição taxonômica completa de cada organismo se encontra disponível anexo 1.

TABELA 2 - DESCRIÇÃO DOS ORGANISMOS SELECIONADOS

Organismo	Acesso NCBI	G+C%	GRAM +/-
Corynebacterium diphtheriae NCTC 13129	NC_002935.2	53,50	+
Corynebacterium glutamicum ATCC 13032	NC_003450.3	53,80	+
Streptococcus agalactiae NEM316	NC_004368.1	35,60	+
Streptococcus mitis B6	NC_013853.1	40,00	+
Streptococcus pyogenes M1 GAS	NC_002737.2	38,50	+
Streptococcus pneumoniae R6	NC_003098.1	39,70	+
Staphylococcus aureus subsp. aureus NCTC 8325	NC_007795.1	32,90	+
Escherichia coli str. K-12 substr. MG1655	NC_000913.3	50,80	-
Escherichia coli CFT073	NC_004431.1	50,50	-
Escherichia coli O157:H7 Sakai	NC_002695.1	50,50	-
Klebsiella pneumoniae subsp. *pneumoniae* HS11286	NC_016845.1	57,50	-
Salmonella enterica subsp. serovar Typhi str. CT18	NC_003198.1	52,10	-
Aeromonas hydrophila subsp. *hydrophila* ATCC 7966	NC_008570.1	61,50	-
Pseudomonas aeruginosa PAO1	NC_002516.2	66,60	-
Vibrio cholerae O1 biovar El Tor N16961 chromosome I	NC_002505.1	47,70	-
Vibrio cholerae O1 biovar El Tor N16961 chromosome II	NC_002506.1	46,90	-

FONTE: O autor (2017).

3.3 FORMATO DE CODIFICAÇÃO DAS INFORMAÇÕES GENÔMICAS

Todas as sequências genômicas dos organismos selecionados foram adquiridas da base de dados NCBI. Foram baixados diversos arquivos com diferentes extensões, referentes aos arquivos necessários para se utilizar nas ferramentas de

predição. Entre eles, .GBK, .FASTA, .FNA, .PTT, .RNT. Todos os dados pertencem a genomas completos.

A relação do número de acessos dos organismos no NCBI, nível de montagem, instituição de envio, data da submissão, arquivos baixados e suas respectivas datas de aquisição, se encontram no anexo 2.

3.4 CONJUNTO DE DADOS PADRÃO OURO

Com base no conjunto de organismos escolhido para nossos testes, buscamos identificar ilhas genômicas que já estavam comprovadas a partir de análises *in vitro* e/ou *in vivo*, a fim de avaliar a sensibilidade e precisão das ferramentas com dados de referência. Pesquisando na literatura, estudos evidenciaram dezesseis ilhas genômicas identificadas em *Escherichia coli* estirpe CFT073 relatadas em três artigos.

3.5 CONTEXTO HISTÓRICO DO ORGANISMO PADRÃO OURO

Em 2007, Lloyd e colaboradores isolaram o organismo a partir de amostra de sangue de um paciente admitido na Universidade de Maryland para tratamento. Com base na sequencia genômica da *Escherichia coli* CFT073, técnicas de PCR e *microarrays* foram utilizadas com análises de hibridização comparativa para estudo do organismo. Para esta pesquisa outros isolados foram usados para comparação contra a *E. coli* CFT073, sendo a *E. coli* K-12 MG1655 utilizado como controle. Às análises revelaram a presença de sete PAIs com seus genes de virulência determinados, três GIs, e três ilhas contendo DNA predominante de bacteriófagos, sendo identificas 13 ilhas (LLOYD; RASKO; MOBLEY, 2007).

Com base nas pesquisas anteriores, Lloyd e colaboradores em 2009, realizaram testes para aferir a capacidade dessas bactérias contendo ilhas colonizar e causar infecção no trato urinário em ratos. Das treze ilhas identificas no estudo anterior, onze foram individualmente eliminadas do genoma e nove ilhas mutantes isogênicas foram testadas (LLOYD et al., 2009). As ilhas mutantes foram construídas a partir de *lambda red recombinase system* (DATSENKO; WANNER, 2000).

No ano de 2011, Vejborg e colaboradores, afim de pesquisar a diversidade dos fatores que levam a causa da infecção urinária, compararam o perfil genômico de 45 estirpes de *E. coli,* juntamente com três estirpes de referência, sendo a *E. coli* CFT073

uma delas. Identificaram as trezes ilhas descritas nas pesquisas anteriores além de outras três ilhas com fatores relacionados com a virulência, contabilizando 16 GIs no total (VEJBORG et al., 2011).

A tabela a seguir (tabela 3) apresenta os dados das ilhas curadas *in vitro* da *E.coli* CFT073 obtidos através dessas pesquisas.

TABELA 3 - DADOS DAS GIs, PAIs E REGIÕES COM DNA DE BACTERIÓFAGOS CURADOS *IN VITRO* DO ORGANISMO DE REFERÊNCIA *Escherichia coli* CFT073

Ilhas	Nome da ilha	Locus Tag$_1$	CDS$_2$	tRNA$_3$	G+C$_4$
1	GI-CFT073-leuX	c5386-c5371	15	leuX	48.15%
2	PAI-CFT073-pheU	c5216-c5143	61	pheU	47.57%
3	GI-CFT073-selC	c4581-c4491	70	selC	47.04%
4	PAI-CFT073-pheV	c3698-c3556	124	pheV	47.08%
5	PAI-CFT073-metV	c3410-c3385	25	metV	53.37%
6	φ-CFT073-smpB	c3206-c3143	49		49.32%
7	GI-CFT073-cobU	c2528-c2482	37		49.68%
8	GI-CFT073-asnW	c2475-c2449	26	asnW	53.12%
9	PAI-CFT073-asnT	c2436-c2418	15	asnT	58.27%
10	PAI-CFT073-serU	c2416-c2392	19	serU	37.65%
11	PAI-CFT073-icdA	c1601-c1518	74		50.23%
12	φ-CFT073-ycfD	c1507-c1481	14		49.78%
13	φ-CFT073-potB	c1475-c1400	51		50.97%
14	PAI-CFT073-serX	c1293-c1165	102	serX	48.76%
15	φ-CFT073-b0847	c0979-c0932	42		50.45%
16	PAI-CFT073-aspV	c0368-c0253	83	aspV	47.43%

FONTE: Adaptado de VEJBORG et al., 2011.
NOTA: $_1$ (Identificadores aplicados a cada gene), $_2$ (Número de sequencias codificadoras), $_3$ (ácido ribonucleico de transferência), $_4$ (Porcentagem de conteúdo guanina e citosina na região), GI (ilhas genômicas), PAI (ilhas de patogenicidade), φ (ilhas contendo predominância de DNA bacteriófagos).

3.6 SÍNTESE DE FUNCIONAMENTO DOS PREDITORES

Primeiramente, para cada ferramenta, foi descrita um breve histórico de sua utilidade e metodologia, bem como seus requisitos de funcionamento. As principais características foram contempladas de maneira detalhada, assimilando com os conceitos biológicos relacionados.

Para avaliarmos as semelhanças e diferenças, envolvendo as estratégias para predição de cada preditor, escolhemos três organismos do conjunto de dados de

interesse, sendo eles *Streptococcus pneumoniae* R6, *Escherichia coli* CFT073 e *Aeromonas hydrophila* ATCC 7966.

3.7 VALIDAÇÃO DAS ILHAS GENÔMICAS

Para estudar e verificar a composição e posição das ilhas genômicas, utilizamos o software Artemis. O Artemis é um software para visualização e anotação de genomas, que disponibiliza várias leituras abrangendo uma mesma região da sequência genômica. Possibilita ainda análise de populações de organismos, variações estruturais e transcriptomas (CARVER, et al., 2012).

Para isso, realizamos um levantamento total das ilhas do padrão ouro, a fim de comparar com todos os preditores, incluindo a contagem de CDSs e composição. A comparação das ilhas em relação a cada preditor foi esclarecido na análise e discussão dos resultados desta pesquisa.

3.8 ANÁLISE E DISCUSSÃO DOS RESULTADOS E CONCLUSÃO

A partir das análises dos dados e da avaliação qualitativa dos preditores, foram realizadas tabelas para auxiliar na interpretação e entendimento do leitor. Como parte da análise quantitativa, utilizamos o *software* Excel para gerar gráficos, a fim de representar visualmente as proporções de cobertura dos preditores em relação às ilhas genômicas dos organismos.

4 RESULTADOS

4.1 ANÁLISE QUALITATIVA DOS PREDITORES – RECURSOS

Todas as ferramentas analisadas possuem uma característica em comum em seus dados de saída, contendo posição inicial e final da ilha predita. Com isto, é possível fazer uso de *softwares* externos para expandir a pesquisa, e se aprofundar no conteúdo contido nas regiões de interesse. No entanto, algumas ferramentas possuem recursos que melhoram a compreensão dos resultados, seja de forma integrada ou ocorrendo um pré-processamento dos dados, para utilizar em outra ferramenta externa.

Alien Hunter, oferece ao usuário a possibilidade de otimizar seus resultados para serem usados em conjunto com a ferramenta de visualização de genomas Artemis. Com isso, é possível identificar as áreas correspondentes como regiões atípicas do genoma, sendo destacadas em cores na visualização dos dados. Devido seu arquivo de entrada ser somente .FASTA, possibilita análises em genomas recém sequenciados.

GI Hunter possui uma ferramenta integrada para visualização dos resultados, *Genomic Island Visualization* - GIV, demonstrando as ilhas identificadas em uma imagem do genoma circular, em suas devidas posições. As extensões de arquivos de entrada dessa ferramenta é um dos problemas que alguns *softwares* estão passando devido atualização do NCBI. Os arquivos .PTT, .RNT, já não estão sendo disponibilizados pelo NCBI. Agora essas duas extensões são fundidas em apenas uma, dificultando o uso da ferramenta devido a necessidade de separar essas informações através de outros programas.

Islandviewer3 possui uma plotagem interativa do genoma, proporcionando ao usuário uma visão ampla de todas as ilhas preditas, com seus devidos produtos e características, indicando genes que possuem relações com fatores de virulência, patogenicidade e resistência a antibióticos. Esta ferramenta é a única que possibilita análises a partir de *drafts* genoma utilizando as mesmas metodologias para predição em genomas completos.

Predict Bias, proporciona uma visão de todo o genoma, com gráficos dos clusters e suas alterações nas composições de sequência. Classifica as ilhas sendo patogênicas ou não patogênicas, e contem opção para análise de comparação de

genomas e geração de árvores filogenéticas. Predict Bias tenta mesclar as duas das principais abordagens para identificar ilhas genômicas. Com a opção para análise comparativa do genoma estudado é possível identificar regiões candidatas a GIs em outros organismos, e pela construção das árvores filogenéticas aferir relações evolutivas entre várias espécies que possam ter um ancestral comum.

Zisland Explorer apresenta uma plotagem da variação do conteúdo G+C ao longo do genoma destacando as regiões candidatas a GIs. Possui arquivos de entrada semelhantes a GI Hunter, contudo, a ferramenta possui opção para se utilizar somente o .FNA, ficando ausente da necessidade de outros softwares para gerar o arquivo .PTT.

GIPSy é a única ferramenta que não apresenta qualquer recurso visual dos resultados, por outro lado, é capaz de classificar as GIs com suas devidas funções.

Esta ferramenta faz uso de várias metodologias integradas para atribuir e categorizar cada GIs identificadas. Devido a necessidade da utilização de um genoma de referência para análise, em conjunto com o seu organismo de estudo, muitas vezes é inviável, pelo fato de haver poucos organismos de referência disponíveis.

Na tabela abaixo (tabela 4), constam algumas informações sobre os preditores, tais como, plataforma necessária para uso, arquivos de entrada e saída. O anexo 3, possui descrição dos *softwares* e seus recursos internos e externos.

TABELA 4 - CARACTERÍSTICAS DESCRITIVA DOS PROGRAMAS PARA PREDIÇÃO DE GI

Preditor	Plataforma	Formato de entrada	Formato de saída
Alien Hunter	Linux	.FASTA	.TXT / .PLOS / .SCO
GI Hunter	Linux	.FNA / .PTT / .RNT	.TXT / PLOT
GIPSy	Linux / Windows	.GBK / .EMBL	.TXT
IslandViewer 3	Web	.GBK / .EMBL	.GBK / .FASTA / .PLOT
Predict Bias	Web	.GBK	.TXT / .PLOT / HTML
Zisland Explorer	Linux / Windows / Web	.FNA / .PTT	.TXT / .PLOT

FONTE: O autor (2017).

As ferramentas de predição de ilhas genômicas possuem muitas características que podem ajudar o usuário em suas pesquisas, como também prejudicar de certa forma no decorrer das análises. Os preditores que são executados por linhas de comando, normalmente exigem mais conhecimento do que aqueles com interface gráfica. Os arquivos de entrada e saída são diversos. Algumas ferramentas como GI

Hunter e Zisland Explorer, fazem uso de algumas extensões de entrada que já não são disponibilizadas separadas pelo NCBI, necessitando de algum programa externa ou script para retirar essas informações.

Para os arquivos de saída, a maioria das ferramentas possuem seus dados em formatos .TXT, facilitando a demonstração e manipulação dos dados. Alguns preditores ainda trabalhando em conjunto com algumas ferramentas externas para visualização de seus resultados em forma de gráficos ou para pesquisa do conteúdo genético das ilhas identificadas, como no caso do Alien Hunter para com o Artemis. Os gráficos de genomas circular e também interativos auxiliaram na compreensão dos resultados, por exemplo, IslandViewer3. Este gráfico do genoma circular interativa possibilita ao usuário uma visão geral de todas as ilhas preditas e os seus produtos que se encontram em cada uma dessas regiões, ajudando nas identificações de processos biológicos importantes.

A interpretação dos resultados varia entre as ferramentas. Algumas possuem mais informações do que somente a posição inicial e final da ilha do genoma, e também predizem a função da ilha. A capacidade de classificar as regiões de acordo com as suas funções é de grande importância para entendimento do comportamento dos organismos. As ferramentas que mostram tanto informações como posições inicias e finais das ilhas, quanto em relação aos seus produtos e funções relacionadas, mostram ao usuário dados que podem ser relevantes para sua pesquisa e entendimento do genoma estudado.

Na tabela abaixo (tabela 5), algumas das vantagens e desvantagens de cada preditor são descritas.

TABELA 5 - VANTAGENS E DESVANTAGENS DOS PREDITORES DE ILHAS GENÔMICAS

Preditores	Vantagens	Desvantagens
Alien Hunter	Trabalha com genomas recém sequenciados, possui *score* de todas as regiões candidatas a GIs em relação ao genoma total.	Não prediz função das ilhas, não detalha os produtos das ilhas, não possui interface, necessita de outro programa para visualizar o conteúdo das ilhas preditas.
GI Hunter	Possui em seus arquivos de saída, uma extensão para se trabalhar diretamente com um visualizador de genoma circular desenvolvido pelo mesmo grupo da ferramenta.	Mostra somente a posição inicial e final da ilha, não prediz função, não detalhe o conteúdo das regiões preditas, necessita de arquivos de entrada com extensão ultrapassadas (.PTT/.RNT).
GIPSy	Prediz a função da ilha, possui informações sobre os produtos presentes nas regiões identificadas, possui interface gráfica.	Necessita de um genoma de referência para rodar em conjunto com o organismo de estudo, exportação dos arquivos de saída possui as informações misturadas.
IslandViewer3	Possui gráfico interativo das ilhas preditas, sendo possível visualizar informações dos produtos que se encontram presentes, exporta os dados das ilhas individualmente (proteínas, genes), possui classificação de genes associados a funções de virulência, patogenicidade, resistência.	Não possibilita ao usuário a escolha do genoma de referência para o método IslandPick, possui alguns resultados diferentes para os mesmos organismos realizando o upload do genoma em comparação com os dados já depositado no banco.
Predict Bias	Prediz função de ilhas de patogenicidade, possui informações de cada um dos produtos das regiões identificadas.	A posição inicial e final da ilha é identificada pela locus tag do primeiro e último gene presente no genoma, e não a posição dos pares de bases.
Zisland Explorer	Possui gráfico do conteúdo G+C das ilhas presentes no decorrer do genoma, informa o tamanho e a quantidade de genes presentes nas GIs candidatas.	Necessita de arquivos de entrada com extensão ultrapassadas (.PTT).

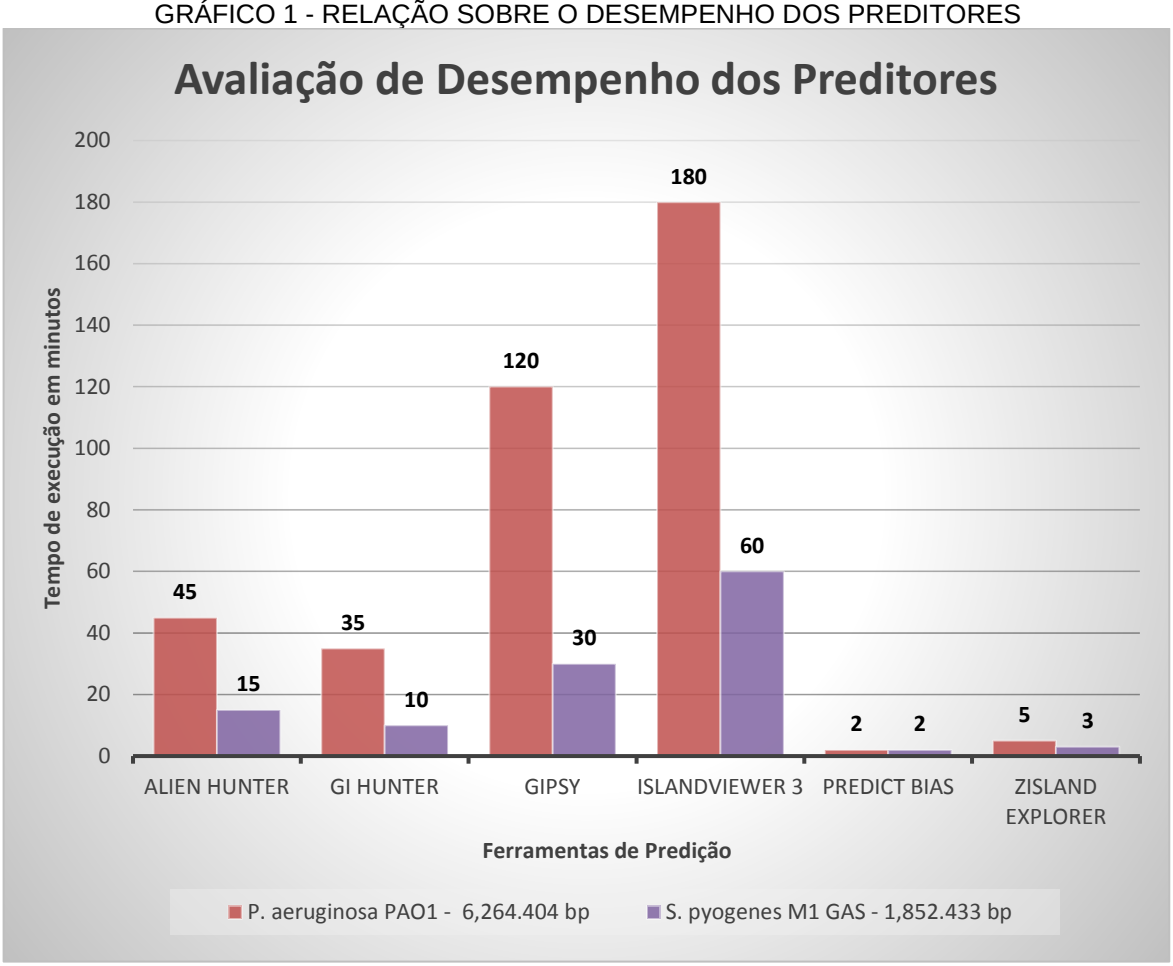
FONTE: O AUTOR (2017).

4.2 AVALIAÇÃO DE DESEMPENHO DOS SOFTWARES

Com o intuito de avaliar o desempenho das ferramentas escolhidas, se tratando do tempo de execução, analisamos dois genomas do nosso grupo de teste, *Pseudomonas aeruginosa* PAO1, contendo cerca de 6,264.404 bp, e *Streptococcus pyogenes* M1 GAS, com 1,852.433 bp. Os dois organismos com mais pares de bases e com menos pares de bases.

O gráfico a seguir (Gráfico 1), demonstra o tempo de execução em minutos de cada ferramenta, comparando os dois genomas.

GRÁFICO 1 - RELAÇÃO SOBRE O DESEMPENHO DOS PREDITORES



FONTE: O AUTOR (2017).

Alien Hunter obteve uma diferença de trinta minutos do maior organismo para o menor, em contrapartida, GI Hunter que faz uso da mesma metodologia integrando Alien Hunter em sua ferramenta, conseguiu diminuir essa diferença para vinte e cinco minutos. GIPSy possui noventa minutos de diferença entre as análises dos genomas, contudo, como este *software* utiliza dois genomas para análise (estudo e referência), seu tempo de execução pode ser variado.

Nas ferramentas *web*, acreditamos que a banda larga não influencia diretamente no tempo que se leva para realizar as análises, entretanto, esta informação não consta nos artigos de publicação ou na página da ferramenta. IslandViewer3 possui a maior diferença de tempo de execução, cerca de cento e vinte minutos entre os dois genomas. Devido a ferramenta passar por diversos processos em suas análises, esse tempo pode ser influenciado pela *query* que se encontra o genoma, em relação a outros organismos que foram enviados anteriormente por outros usuários.

Diferente do Predict Bias, seu tempo não foi influenciado pelo tamanho dos genomas. Devido esta ferramenta, ser de certa forma, um banco de dados em conjunto, pode-se esperar que algumas anotações já poderiam estar pré-processadas. Zisland Explorer, apresentou uma pequena diferença, mas também não foi muito influenciada pelo tamanho dos organismos analisados. Em resumo, todas as ferramentas apresentaram tempo de execução relativamente pequeno, e nenhuma apresentou erros durante a execução.

4.3 COMPARAÇÃO DOS RESULTADOS DOS PREDITORES COM O PADRÃO OURO

Com o conjunto de dados estabelecido como “Padrão Ouro”, do organismo *Escherichia coli* CFT073, derivado das pesquisas relatadas nos três artigos, (LLOYD; RASKO; MOBLEY, 2007), (DATSENKO; WANNER, 2000). (VEJBORG et al., 2011), buscamos identificar qual ferramenta conseguiu realizar as predições mais próximas das dezesseis ilhas curadas *in vivo*.

Identificamos no primeiro momento a similaridade entre as ilhas curadas e das preditas pelas ferramentas, com base no tamanho das ilhas do padrão ouro, a partir de porcentagens de acertos entre 25%, 50% e 75% entre início e final da ilha pela posição no genoma. Essas porcentagens devem constar a região da ilha curada sendo determinada pela região da ilha predita. Na tabela abaixo (tabela 6), retrata o número de acertos dos preditores de acordo com a porcentagem de cobertura das ilhas padrão ouro.

TABELA 6 - ILHAS PREDITAS PELAS FERRAMENTAS *VERSUS* 16 ILHAS CURADAS *IN VITRO ESCHERICHIA COLI* CFT073

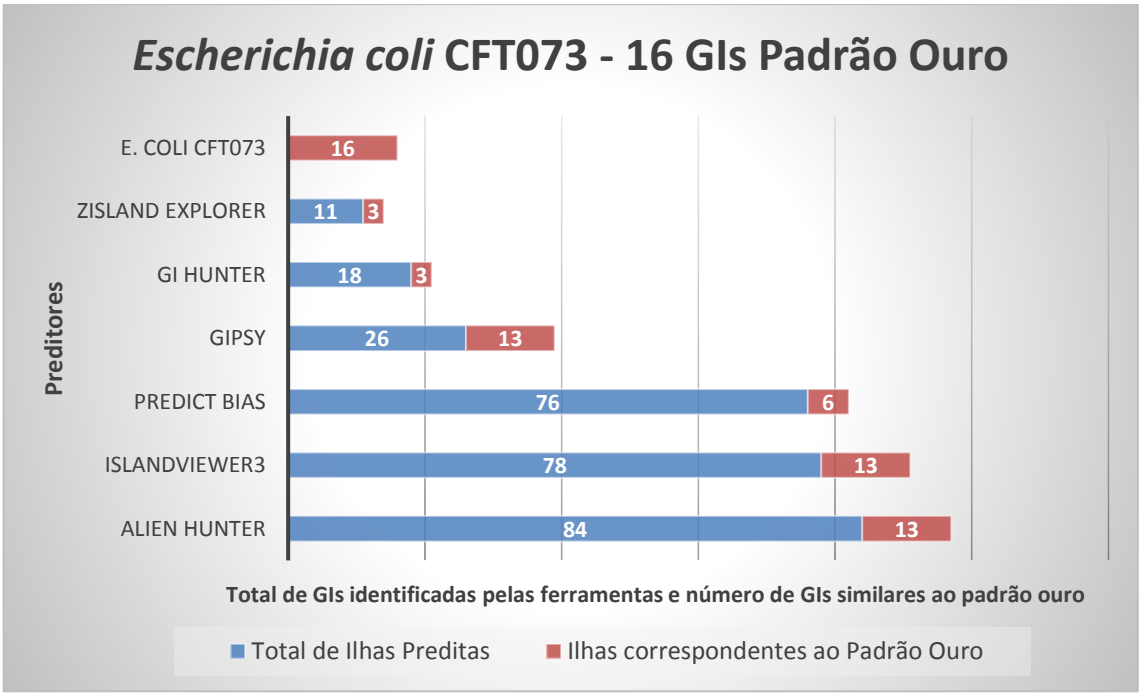
Preditores	25%₁	50%₂	75%₃
Zisland Explorer	3	3	3
GI Hunter	8	5	3
Predict Bias	9	7	6
IslandViewer3	14	14	13
GIPSy	15	14	13
Alien Hunter	15	14	13

FONTE: O autor (2017).

NOTA: $_1$, $_2$, $_3$ (Porcentagem em relação ao tamanho da ilha, posição inicial e final e sua quantidade de acertos por cada preditor).

O gráfico a seguir (gráfico 2) representa o total de ilhas identificadas pelas ferramentas e a quantidade de ilhas pertencentes ao padrão ouro de acordo com a porcentagem de 75% de acerto.

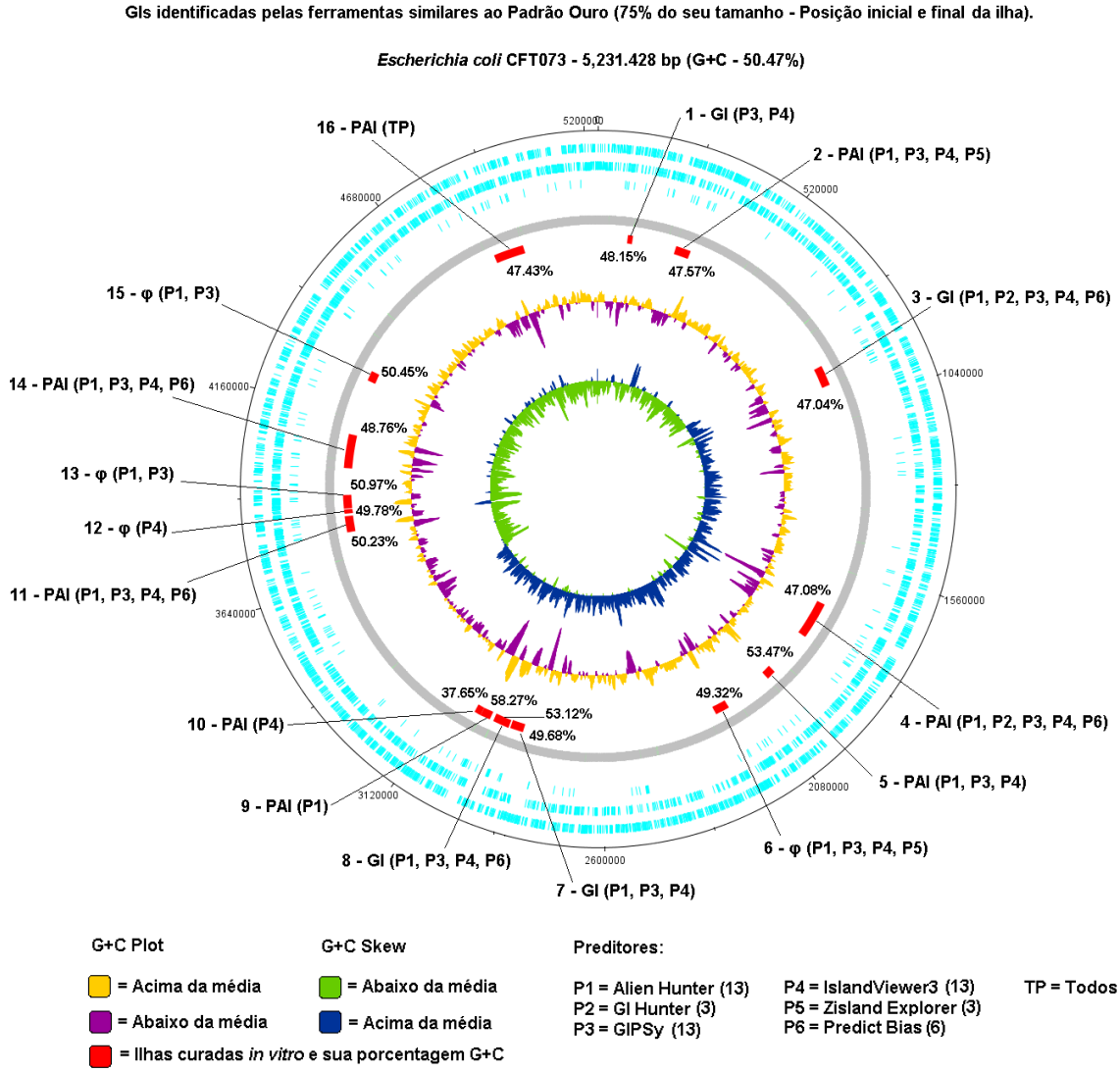
GRÁFICO 2 - PREDIÇÃO TOTAL DAS ILHAS GENÔMICAS EM ESCHERICHIA COLI CFT073 E SUAS REGIÕES SIMILARES AS ILHAS CURADAS *IN VITRO*



FONTE: O AUTOR (2017).

Em um segundo momento, com base nos resultados obtidos, buscamos saber quais desses preditores identificaram ilhas iguais e diferentes, com base na porcentagem de 75%. Na figura do genoma circular (figura 9), criado a partir do DNAPlotter da ferramenta Artemis, é possível visualizar as dezesseis ilhas curadas *in vitro* e sua disposição no genoma, assim como, informações de conteúdo G+C% do genoma inteiro e de cada ilha, e quais ferramentas conseguiram identificá-las.

FIGURA 9. REPRESENTAÇÃO DO GENOMA CIRCULAR DE *ESCHERICHIA COLI* CFT073



FONTE: Adaptado do DNAPlotter da ferramenta Artemis, CARVER, et al., 2012.
NOTA: Regiões das Ilhas genômicas identificadas *in vitro,* analisadas a partir do software Artemis. GI: ilhas genômicas; PAI: ilhas de patogenicidade; φ: ilhas contendo DNA predominante de fagos.

Das dezesseis ilhas descritas no padrão ouro, três regiões (19%) foram identificadas apenas por uma ferramenta, (ilha 9) Alien Hunter, (ilha 10) IslandViewer3, (ilha 12) IslandViewer3. Somente uma única ilha (ilha 16) em que todos os preditores conseguiram assimilar.

A primeira ilha representada pelo número 1, é uma ilha genômica, contendo uma variação de G+C (48.15%), com tRNA associado (leuX), e possuindo uma integrasse sendo ausente de transposases. A ferramenta GIPSy obteve resultado da região predita com conteúdo G+C de 46.07%, conseguiu associar o gene de tRNA e a integrase, contudo sua variação de G+C é resultante de aquisição de CDS a mais

do que a ilha do padrão ouro. IslandViewer3 também conseguiu identificar esta região. Obteve um conteúdo G+C de 48.77%, não conseguindo associar o gene de tRNA e a integrasse, podendo ter sua variação de G+C derivada pela perda de CDS em sua predição.

A segunda ilha representada pelo número 2, é uma ilha de patogenicidade, contendo uma variação de G+C (47.57%), com tRNA associado (pheU), e possuindo 3 integrase e 11 transposases. Entre as 4 ferramentas que identificaram esta região, GIPSy e Alien Hunter obtiveram resultados mais significativos. GIPSy apresentou uma variação de G+C de (47.44%), associou o gene de tRNA, juntamente como as integrases e transposases presentes. Alien Hunter identificou uma variação de G+C de (47.58%), associando também o gene de tRNA e os genes de mobilidade.

A terceira ilha representada pelo número 3, é uma ilha genômica, contendo uma variação de G+C (47.04%), com tRNA associado (selC), e possuindo 2 integrase e 10 transposases. Entre as 5 ferramentas que identificaram esta região, GIPSy, Alien Hunter e IslandViewer3 tiveram melhores resultados. GIPSy apresentou uma variação de G+C de (47.29%), associou o gene de tRNA, juntamente como as integrases e transposases presentes. Alien Hunter identificou uma variação de G+C de (47.20%), associando também o gene de tRNA e os genes de mobilidade juntamente com IslandViewer3, possuindo a mesma variação de conteúdo G+C%, identificando os mesmos produtos.

A quarta ilha representada pelo número 4, é uma ilha de patogenicidade, contendo uma variação de G+C (47.08%), com tRNA associado (pheV), e possuindo 3 integrase e 20 transposases. Entre as 5 ferramentas que identificaram esta região, novamente, GIPSy, Alien Hunter e IslandViewer3 tiveram melhores resultados. GIPSy identificou a região com variação de conteúdo G+C de (47.18%), juntamente com todos os genes de mobilidade. Alien Hunter e IslandViewer3 obtiveram uma conteúdo G+C de (47.00%) e (46.98%) respectivamente, não conseguindo associar o gene de tRNA em suas predições e também 1 integrase que se encontrava presente.

A quinta ilha representada pelo número 5, é uma ilha de patogenicidade, contendo uma variação de G+C (53.37%), com tRNA associado (pheV), e não possuindo integrase e transposases. Entre as 3 ferramentas que identificaram esta região, GIPSy, Alien Hunter tiveram melhores resultados. GIPSy apresentou uma variação de G+C de (52.89%), associando o gene de tRNA presente e mais 2 tRNAs que se encontravam após o tRNA designado para ilha, podendo indicar o motivo da

variação de G+C% se encontrar um pouco a baixo da ilha curada. Alien Hunter identificou uma variação de G+C de (53.48%), se aproximando mais do conteúdo G+C da ilha curada, mas não associou o gene de tRNA.

A sexta ilha representada pelo número 6, é uma ilha com grande presença de DNA bacteriófagos, contendo uma variação de G+C (49.32%), gene de tRNA ausente, e possuindo 1 integrase e 1 transposases. Entre as 4 ferramentas que identificaram esta região, Alien Hunter e IslandViewer3 tiveram melhores resultados. Alien Hunter identificou esta região com conteúdo G+C de (48.99%) e IslandViewer3 com (49.15%). Ambas as ferramentas assimilaram os genes de mobilidade presentes na ilha.

A sétima ilha representada pelo número 7, é uma ilha genômica, contendo uma variação de G+C (49.68%), gene de tRNA ausente, e possuindo 1 integrase e 7 transposases. Entre as 3 ferramentas que identificaram esta região, GIPSy obteve melhores resultados. Seu conteúdo G+C é de (49.42%), identificando todos os genes de interesse da ilha.

A oitava ilha representada pelo número 8, é uma ilha genômica, contendo uma variação de G+C (53.12%), com tRNA associado (asnW), e possuindo 1 integrase e 2 transposases. Entre as 4 ferramentas que identificaram esta região, GIPSy e IslandViewer3 tiveram melhores resultados. GIPSy obteve conteúdo G+C de (53.03%) e IslandViewer3 (53.38%). Somente GIPSy conseguiu identificar o tRNA associado e todos os genes de mobilidade, já IslandViewer3 não identificou o tRNA juntamente com a integrase presente na ilha.

A nona ilha representada pelo número 9, foi identificada apenas pela ferramenta Alien Hunter. É uma ilha de patogenicidade devido à presença do gene *fyuA*. Este gene promove a codificação para o receptor de *yersiniabactin*, um sideróforo encontrada em bactérias patogênicas. *FyuA*, possui grande importância na formação de biofilmes em ambientes que são desfavorecidos da presença de ferro, como por exemplo, a urina humana (HANCOK; FERRIERES; KLEMM, 2008). Esta ilha contém quatorze CDS no total, sendo flanqueado pelo gene de tRNA (asnT) seguido por uma integrase. A transposase se encontra no meio da ilha e *fyuA* em sua extremidade. Alien Hunter não conseguiu identificar o gene de tRNA, mas assimilou a presença do gene *fyuA* em sua predição. Em seu limiar de corte geral para identificar as regiões atípicas do genoma consta um *Threshold* de 11.44, e seu *score* atribuído é de 18,24.

A décima ilha representada pelo número 10, foi predita apenas por IslandViewer3. Também é uma ilha de patogenicidade pela presença do gene *tcpC*, responsável por interferir na resposta imune inata do hospedeiro (ERJAVEC et al., 2010). Contendo vinte e seis CDS em sua região, sendo flanqueado pelo gene de tRNA (serU), e uma integrase em sua outra extremidade, sendo ausente de transposase. O gene *tcpC*, se encontra no meio da ilha. No entanto, na anotação do .GBK, este gene se encontra marcado como uma proteína hipotética. De acordo com UNIPROT, o resultado derivado do BLAST deste produto em relação ao gene *tcpC*, possui uma identidade de 100% com uma *Query Length* de 207 e *Match Length* de 307. IslandViewer3 conseguiu identificar toda a região e suas CDS.

A décima primeira ilha representada pelo número 11, é uma ilha patogênica, contendo uma variação de G+C (50.23%), gene de tRNA ausente, e possuindo 2 integrase e 4 transposases. Entre as 4 ferramentas que identificaram esta região, GIPSy e IslandViewer3 tiveram melhores resultados. GIPSy obteve um conteúdo G+C de (50.02%) em sua predição e IslandViewer3 (48.97%). GIPSy conseguiu identificar todos os genes de mobilidade, e IslandViewer3 associou 1 integrase e 2 transposases.

A décima segunda ilha, identificada apenas por IslandViewer3, é uma região que contém predominância de DNA proveniente dos bacteriófagos. Esta ilha não contem genes de tRNA e somente uma integrase.

A décima terceira ilha representada pelo número 13, é uma ilha com grande presença de DNA bacteriófagos, contendo uma variação de G+C (50.97%), gene de tRNA se encontra ausente no artigo de referência das ilhas curadas *in vitro*, mas em nossas análises identificamos 3 tRNAs nesta região, possui ainda 1 integrase e 1 transposases. Duas ferramentas identificaram esta região, GIPSy e Alien Hunter. GIPSy obteve uma variação de conteúdo G+C de (51.66%) e Alien Hunter (52.19%). Ambas as ferramentas identificam os 3 genes de tRNA que se encontravam no meio da ilha juntamente com a transposase, ficando ausente a identificação da integrasse nas duas ferramentas.

A décima quarta ilha representada pelo número 14, é uma ilha de patogenicidade, contendo uma variação de G+C (48.76%), com tRNA associado (serX), e possuindo 3 integrase e 12 transposases. Entre as 4 ferramentas que identificaram esta região, GIPSy, Alien Hunter e IslandViewer3 tiveram resultados melhores. O conteúdo G+C da ilha predita por GIPSy foi de (48.73%), Alien Hunter (48.43%), IslandViewer3 (48.45%). Somente GIPSy conseguiu identificar todos os

genes de interesse, e ambas as ferramentas Alien Hunter e IslandViewer3 não associaram o gene de tRNA e 1 transposase, entretanto todas as integrases foram identificadas.

A décima quinta ilha representada pelo número 15, é uma ilha com grande presença de DNA bacteriófagos, contendo uma variação de G+C (50.45%), gene de tRNA ausente, e possuindo 1 integrase e nenhuma transposases. Duas ferramentas tiveram resultados satisfatórios nesta região, GIPSy e Alien Hunter. O conteúdo de variação de G+C predito por GIPSy foi de (50.28%) e Alien Hunter (50.47%). Ambas as ferramentas identificam o gene integrasse presente na ilha.

Entre todas as ilhas do padrão ouro, somente a décima sexta ilha, representada pelo número 16, foi identificada por todos os preditores. Esta região é caracteriza como uma ilha de patogenicidade, contendo cinco genes relacionados a fatores de virulência, entre eles*, fpbABC , cdiA, picU, tosCBDA, vat* (VEJBORG et al., 2011). Não possui integrase em sua composição e 43% da ilha é composta por proteínas hipotéticas e não caracterizadas. Um resumo do conteúdo das predições dessa ilha se encontra na tabela abaixo (tabela 7).

TABELA 7 - CARACTERÍSTICA DA DÉCIMA SEXTA ILHA DO PADRÃO OURO ENCONTRADAPOR TODOS OS PREDITORES

P. Ouro$_1$ /Preditores	tRNA$_2$	Transposase	G+C%$_3$	Genes Vir.$_4$	Prot Hip$_5$	Não Carac$_6$	Total CDS$_7$
PAI - 16	aspV	12	47.43%	5	38	5	100
IslandViewer3	presente	12	47.43%	5	38	5	100
GIPSy	presente	12	47.38%	5	38	5	99
Alien Hunter	ausente	8	46.97%	4	33	5	84
Zisland Explorer	ausente	8	46.35%	4	35	4	81
GI Hunter	ausente	8	46.34%	3	32	4	79
Predict Bias	ausente	7	46.30%	3	31	4	77

FONTE: O Autor (2017).
NOTA: Em destaque, décima sexta ilha do padrão ouro *Escherichia coli* CFT073. $_1$ (Padrão Ouro), $_2$ (ácido ribonucleico de transferência), $_3$ (Porcentagem de conteúdo guanina e citosina na região), $_4$ (Genes de Virulência), $_5$ (Proteínas Hipotéticas), $_6$ (Proteinas não caracterizadas), $_7$ (Número de sequencias codificadoras).

Com todas as ilhas identificadas, o conteúdo genético de cada uma é levantado de forma geral, para se obter uma visão ampla das possíveis perdas de informações biológicas, tendo como referência os produtos contidos nas ilhas curadas *in vitro*. Entretanto, na comparação dos dados relatados nos artigos com os encontrados na ferramenta de visualização de genomas Artemis com o arquivo de anotação proveniente do NCBI, percebemos que ocorrem algumas diferenças, como no número

total de CDS presentes em cada ilha. Há também a presença de três tRNAs na décima terceira ilha e um tRNA a mais na décima quarta ilha, que se encontravam ausentes nas informações contidas nos artigos. Na tabela abaixo (tabela 8), é possível visualizar essas particularidades.

TABELA 8 - DADOS DAS ILHAS DESCRITAS NO PADRÃO OURO EM RELAÇÃO AS REGIÕES ENCONTRADAS NAS ANOTAÇÕES FORNECIDAS PELO NCBI

Ilhas	Locus Tag$_1$	/NCBI	CDS$_2$	/NCBI	tRNA$_3$	/NCBI	G+C$_4$	/NCBI
1-GI	c5386-c5371	c5387-c5371	15	13	leuX	leuX	48.15%	48.15%
2-PAI	c5216-c5143	c5216-c5143	61	62	pheU	pheU	47.57%	47.57%
3-GI	c4581-c4491	c4581-c4491	70	77	selC	selC	47.04%	47.04%
4-PAI	c3698-c3556	c3698-c3556	124	131	pheV	pheV	47.08%	47.08%
5-PAI	c3410-c3385	c3411-c3385	25	25	metV	metV	53.37%	53.37%
6-φ	c3206-c3143	c3206-c3143	49	62			49.32%	49.32%
7-GI	c2528-c2482	c2529-c2482	37	47			49.68%	49.68%
8-GI	c2475-c2449	c2475-c2449	26	24	asnW	asnW	53.12%	53.12%
9-PAI	c2436-c2418	c2436-c2418	15	14	asnT	asnT	58.27%	58.27%
10-PAI	c2416-c2392	c2416-c2392	19	26	serU	serU	37.65%	37.65%
11-PAI	c1601-c1518	c1601-c1519	74	84			50.23%	50.23%
12-φ	c1507-c1481	c1508-c1480	14	21			49.78%	49.78%
13-φ	c1475-c1400	c1475-c1400	51	67			50.97%	50.97%
14-PAI	c1293-c1165	c1292-c1165	102	118	serX	serX	48.76%	48.76%
15-φ	c0979-c0932	c0979-c0932	42	43			50.45%	50.45%
16-PAI	c0368-c0253	c0369-C_RS0120	83	100	aspV	aspV	47.43%	47.43%

FONTE: Adaptado de (VEJBORG et al., 2011); O Autor (2017).
NOTA: $_1$ (Identificadores aplicados a cada gene), $_2$ (Número de sequencias codificadoras), $_3$ (ácido ribonucleico de transferência), $_4$ (Porcentagem de conteúdo guanina e citosina na região). GI: ilha genômica; PAI: ilha de patogenicidade; φ: ilha contendo DNA predominante de fagos.

Apesar das diferenças do número total de CDS presentes nas ilhas do padrão ouro em relação as ilhas identificadas na ferramenta Artemis, o seu conteúdo G+C% não teve alteração. Uma das possíveis causas dessas diferenças, podem ser devido aos artigos não enunciarem como essa contagem de CDS foi realizada, se é incluído proteínas hipotéticas, proteína não caracterizadas, ou até mesmo pseudo genes, por exemplo. O gene de tRNA sempre é encontrado flanqueando a ilha após o último produto, não sendo englobado na região. Portando, sabemos que o tRNA não faz parte da contagem total de CDS.

Para avaliarmos o conteúdo das ilhas encontradas pelos preditores, nos baseamos a partir dos achados pela ferramenta Artemis. Realizamos um

levantamento total das dezesseis ilhas do padrão ouro com os principais produtos presentes, como, presença de tRNA, integrase, transposase, proteínas hipotéticas e não caracterizadas, e número de CDS contidos na região, afim de compararmos com os resultados totais dos preditores.

Incluímos todos os produtos na contagem de CDS, menos o gene de tRNA. Para atribuirmos se determinada ferramenta conseguiu identificar este gene, deverá englobar o mesmo em sua região ou realizar a sua predição até último produto antes do tRNA. Nenhuma ferramenta apresentou predições exatas de posições iniciais e finais comparadas com o padrão ouro, resultando em perdas de CDS ou até mesmo aquisição de material genético. Para que isso não afete na contagem total dos genes, toda ilha identificada pelos preditores que conter CDS a mais de qualquer produto avaliado, em relação as ilhas do padrão ouro, este conteúdo extra será excluído da contagem final. Na tabela abaixo (tabela 9), é demonstrado os resultados obtidos de todas as ilhas identificadas pelas ferramentas em relação aos produtos das dezesseis ilhas do padrão ouro.

TABELA 9 - TOTAL DE CDS PRESENTES NAS 16 ILHAS DESCRITAS NO PADRÃO OURO EM COMPARAÇÃO COM OS RESULTADOS DOS PREDITORES

P. Ouro[1] /Preditores	tRNA[2]	Integrase	Transposase	Prot Hip[3]	Não Carac[4]	Pseudo[5]	Total CDS[6]	Cobertura %
E.coli CFT073	13	22	81	309	51	70	914	100%
GIPSy	11	16	80	276	40	68	832	91%
Alien Hunter	5	13	70	241	37	53	748	81%
IslandVIewer3	3	12	75	239	37	57	720	78%
Predict Bias	1	4	26	101	17	25	290	22%
GI Hunter	0	0	15	56	11	14	158	17%
Zisland Explorer	0	1	13	47	9	12	149	16%

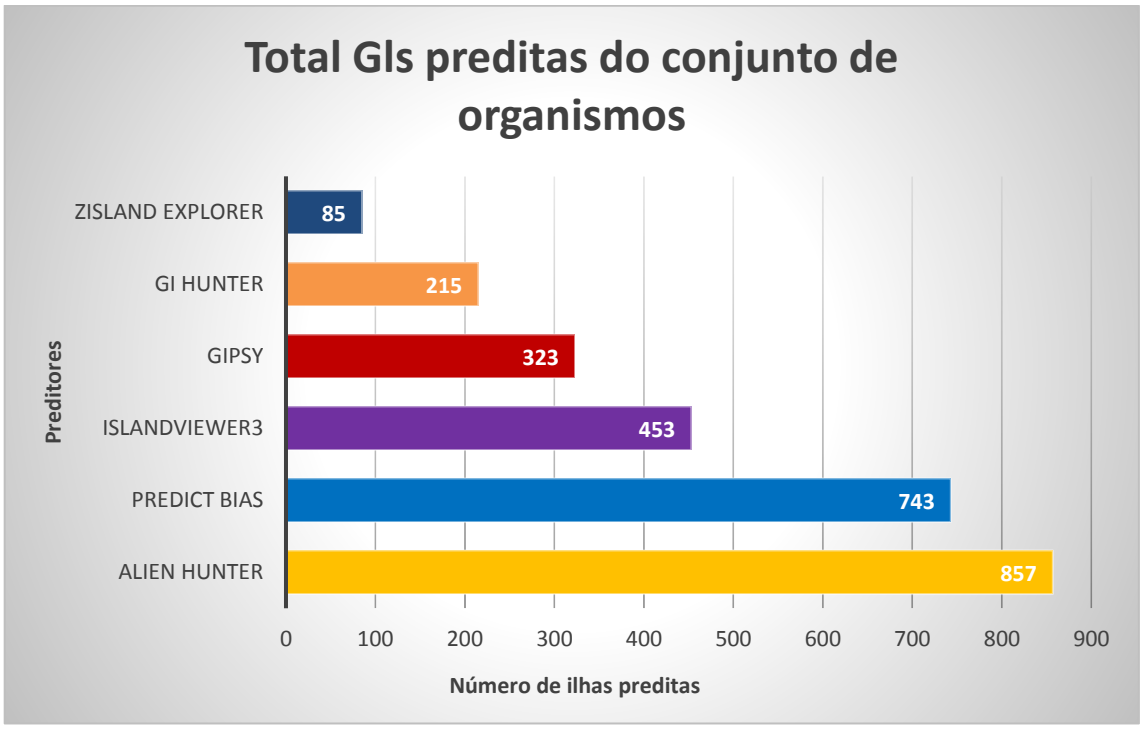
FONTE: O autor (2017).
NOTA: Em destaque, total de produtos presentes nas 16 ilhas do Padrão Ouro. [1] (padrão ouro), [2] (ácido ribonucleico de transferência), [3] (proteínas hipotéticas), [4] (proteínas não caracterizadas), [5] (pseudogenes), [6] (Número de sequencias codificadoras).

Observamos que os preditores Zisland Explorer, GI Hunter, Predict Bias perderam muitos produtos, tanto no geral de CDS quanto genes importantes e muito característico das GIs, como integrases, transposases e tRNAs. Diferente das outras ferramentas. IslandViewer3 e Alien Hunter não identificam muitos genes de tRNAs, mas a sua predição total conseguiu cobrir boa parte das CDS. GIPSy apresentou bom resultado, conseguindo identificar a maioria dos produtos.

4.4 DESEMPENHO DOS PREDITORES NO CONJUNTO TOTAL DE ORGANISMOS

Todas as ferramentas avaliadas possuem o mesmo objetivo, contudo, estratégias diferentes são utilizadas em cada uma, afetando diretamente o resultado final. As abordagens de composição de sequência e comparação genômica são as mais utilizadas entre os softwares de predição de ilhas genômicas. Existem programas que tentam mesclar esses dois métodos, para tentar minimizar resultados falsos positivos e negativos. No gráfico abaixo (gráfico 3), são apresentados os dados em relação ao total de ilhas preditas por cada ferramenta em nosso conjunto de organismos. O número de ilhas identificadas em cada organismo separado se encontra no anexo 4.

GRÁFICO 3 – TOTAL DE GIs PREDITAS NO CONJUNTO DE ORGANISMOS



FONTE: O autor (2017).
NOTA: Total de 15 organismos testados entre todos os preditores.

Podemos observar, que, mesmo ferramentas que contenham metodologias integradas, como no caso do GI Hunter para com Alien Hunter, seus resultados podem ser diferentes. Já na comparação com Predict Bias, Alien Hunter possui quase o mesmo número de ilhas preditas. Essas duas ferramentas apresentaram mais

sensibilidade durante suas análises. IslandViewer3 é uma das ferramentas que busca integrar os métodos de predição, tanto voltados para composição de sequência, quanto para comparação genômica, possuindo número total de predições em torno da metade dos outros preditores.

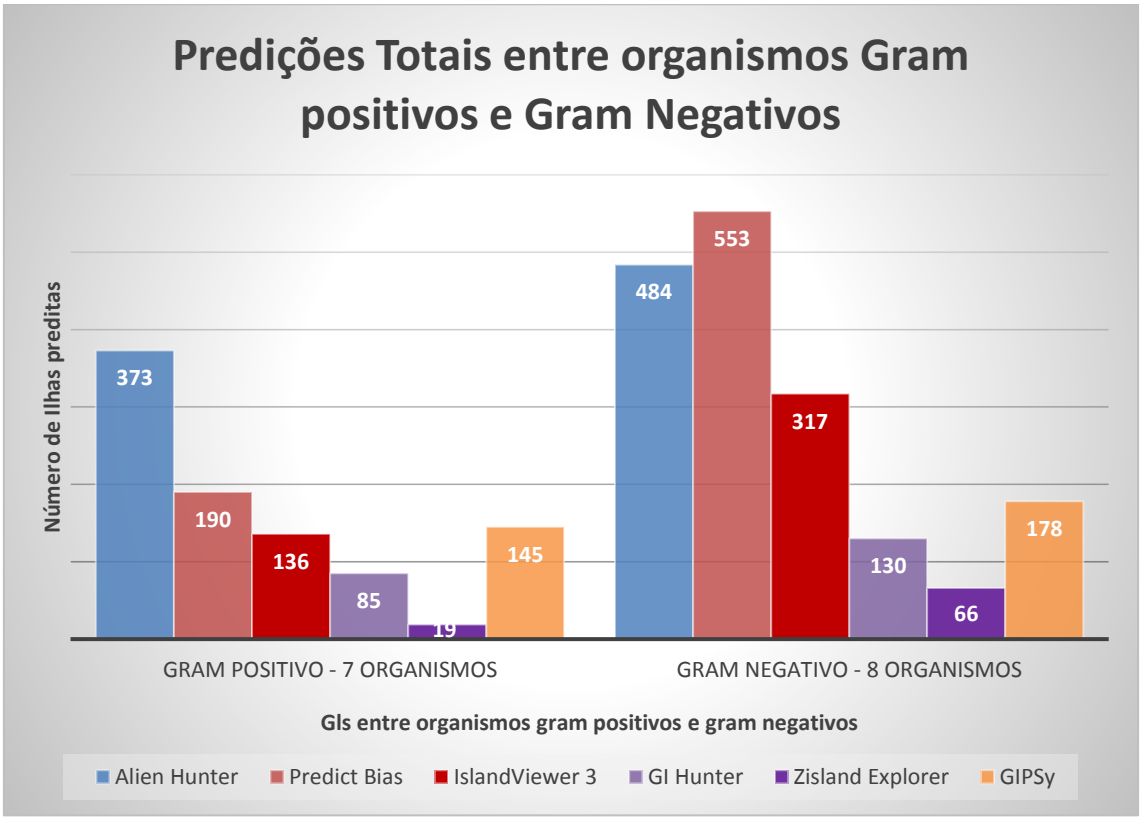
GIPSy possui uma característica incomum. Devido a necessidade da ferramenta requerer um genoma para estudo e outro para referência, seus resultados totais podem variar de acordo com o genoma escolhido. Como comparado anteriormente, GI Hunter utiliza a mesma metodologia que Alien Hunter e mesmo assim não alcançou a metade do total de predições. Isso pode ter ocorrido devido a outros recursos que GI Hunter utilizou para realizar a suas análises. Na construção do modelo de predição, esta ferramenta fez uso de determinado conjunto de dados para treinamento, sendo afetado diretamente pela qualidade desses dados.

Zisland Explorer, faz uso de método bem especifico voltado para variação de conteúdo G+C, contudo, se determinado organismo não ter uma diferença muito grande deste conteúdo, pode ser que algumas ilhas passem despercebidas, tendo um número muito pequeno em relação ao total de ilhas preditas em todos os organismos.

4.5 TOTAL DE GIS PREDITAS – GRAM POSITIVOS E NEGATIVOS

Separamos o número total de ilhas preditas entre as espécies gram positivas (sete organismos), e gram negativa (oito organismos), afim de avaliarmos se as ferramentas e metodologias, integradas ou desenvolvidas, proporcionavam resultados muito diferentes. No gráfico abaixo (gráfico 4), é possível visualizar a diferença dos resultados totais entre as predições dos organismos gram positivos e negativos. Os resultados de cada organismo separado se encontram no anexo 5.

GRÁFICO 4 - TOTAL DE GIs PREDITAS EM ORGANISMOS GRAM POSITIVOS E NEGATIVOS DO CONJUNTO DE ORGANISMOS TESTE



FONTE: O autor (2017).
NOTA: Total de pares de bases entre organismos gram positivos (16.868.541 bp), gram negativos (40.556.825 bp). FONTE: O autor (2017).

Apesar de organismos gram negativos possuem genomas relativamente maiores quando comparados aos gram positivos, em nosso conjunto de dados, três preditores apresentaram resultados similares ao número total de predições quanto a diferença entre os organismos.

Alien Hunter, GI Hunter e GIPSy, identificaram quase o mesmo número de ilhas entre os organismos gram positivos e gram negativos. E Predict Bias, IslandViewer3, Zisland Explorer, mostraram grandes diferenças no número total de ilhas preditas entre os organismos. No anexo 6, se encontra a descrição do genoma de cada organismo, contendo seu tamanho, conteúdo G+C%, total de genes, total de CDS, rRNAs, tRNAs, integrases, transposases, proteínas hipotéticas e pseudo genes.

4.6 ANÁLISE DE SIMILARIDADE DE RESULTADOS ENTRE AS FERRAMENTAS

Para avaliarmos os resultados entre as ferramentas, se tratando de suas semelhanças e diferenças, envolvendo todas as suas estratégias para predição, escolhemos três organismos do nosso conjunto de dados. *Escherichia coli* CFT073, por ser o nosso organismo padrão ouro sendo agora analisado o genoma completo do organismo e não somente as 16 ilhas curadas *in vitro*; *Aeromonas hydrophila* subsp. hydrophila ATCC 7966, devido à baixa presença de genes relacionados com mobilidade e um conteúdo G+C% de 61.50%, acima da *E.coli* CFT073; e um organismo gram positivo *Streptococcus pneumoniae* R6, com genes integrase limitados ao longo do seu genoma, mas com transposases presentes, e em razão de possuir uma porcentagem G+C% menores que os outros dois organismos, com 39.70%.

Os resultados foram ponderados da seguinte maneira. Primeiro, optamos por trabalhar somente com a posição inicial e final de cada ilha. Com esses dados, montamos uma intersecção entre as predições de cada ferramenta, somamos todos as predições das 6 ferramentas para cada organismo, com o intuito de avaliar cada preditor separadamente comparando os seus resultados com as demais predições. Devido à falta de dados de referência, não sabemos quais predições assumem resultados falsos positivos e falsos negativos, então a busca por regiões de similaridade foi feita a partir de 50% entre o tamanho da ilha. Segundo, aferimos as regiões similares entre as ferramentas e diminuímos do número total de predições, com isso, obtivemos regiões totais para compararmos com as predições similares de cada ferramenta. Identificamos também ilhas que não foram preditas por nenhuma outra ferramenta.

De acordo com (HACKER; KEPPER, 2000) as ilhas com mais de 10 kb são classificadas como ilhas genômicas, e menores sendo ilhotas genômicas. Diante disto, extraímos algumas informações, como, o tamanho médio das ilhas identificadas por cada preditor e a relação entre eles, as regiões com menores e maiores composições seguido da contagem de genes presentes, e o número de ilhas que podem ser designadas como ilhas genômicas ou ilhotas.

4.6.1 *Escherichia coli* CFT073

As ilhas presentes ao longo do genoma, possuem em média um tamanho de 30 kb, a maior região predita contém 192 genes dentro de 213 kb, e a menor, 4 genes em 3 kb. O número total de ilhas preditas pelas ferramentas é de 304, sendo 156 designadas como ilhas genômicas, e 148 como ilhotas. A tabela abaixo (tabela 10), demonstra o resumo dessas informações.

TABELA 10 - INFORMAÇÕES DAS ILHAS EM *ESCHERICHIA COLI* CFT073

Preditores	Média[1]	Mínima[2]	<Genes[3]	Máxima[4]	>Genes[5]	Total Pred[6]	<10 kb[7]	>10 kb[8]
GI Hunter	57749	5354	6	213265	192	18	3	15
Predict Bias	27670	3120	4	188170	180	76	22	54
GIPSy	28540	5751	8	113961	118	38	16	22
IslandViewer3	10878	4009	8	102064	115	78	58	20
Zisland Explorer	45354	5477	7	94941	90	11	2	9
Alien Hunter	13053	4027	5	39550	40	83	47	36

FONTE: O Autor (2017).
NOTA: [1] (Tamanho médio das ilhas preditas por cada ferramenta), [2] (Tamanho mínimo das ilhas preditas por cada ferramenta), [3] (Quantidade de genes presentes nas ilhas de tamanho mínimo por cada ferramenta), [4] (Tamanho máximo das ilhas preditas por cada ferramenta), [5] (Quantidade de genes presentes nas ilhas de tamanho máximo por cada ferramenta), [6] (Total de ilhas preditas por cada ferramenta), [7] (Quantidade total de ilhotas preditas por cada ferramenta), [8] (Quantidade total de GIs preditas por cada ferramenta).

O gráfico para visualização dos resultados de similaridades entre as predições foi elaborado de acordo com a tabela a seguir. Optamos por demostrar esses resultados na tabela somente pela ferramenta que conseguiu melhor desempenho. Os resultados dos outros preditores foram elaborados da mesma forma de acordo com a tabela abaixo (tabela 11).

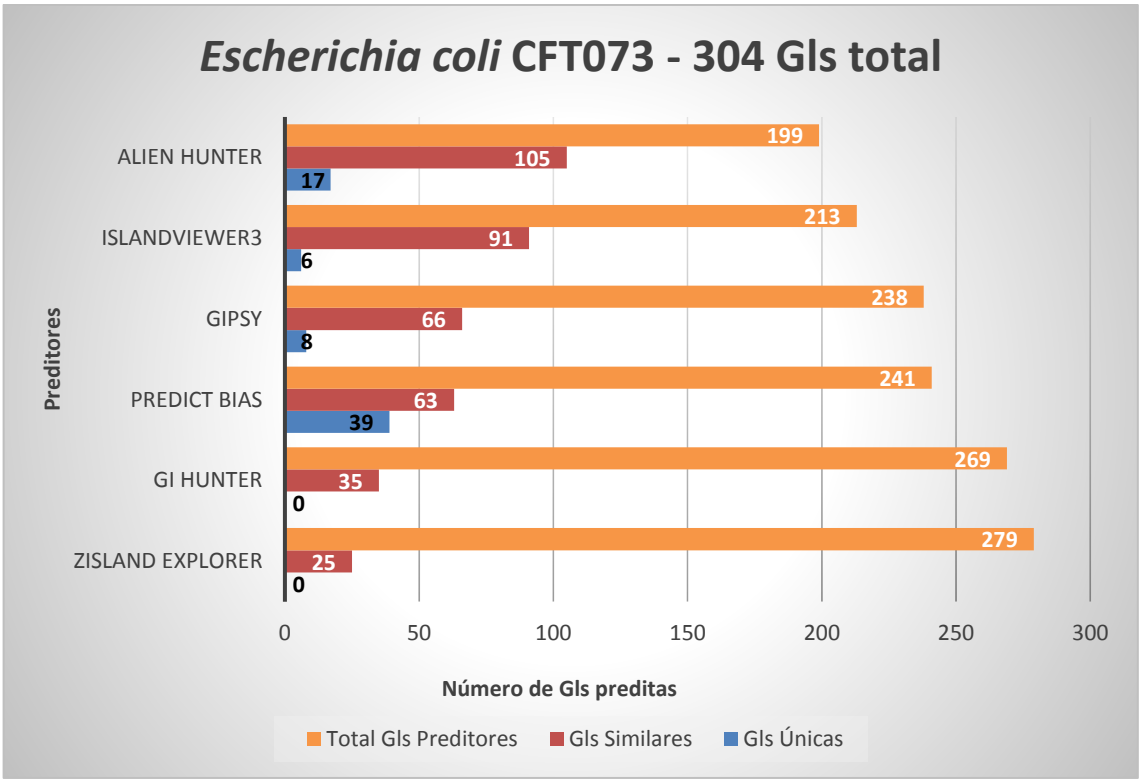
TABELA 11 - PREDITOR COM MELHOR DESEMPENHO EM *ESCHERICHIA COLI* CFT073

Preditores	Total predição[1]	Similares[2]	T. predição s/ similares[3]	GIs Únicas[4]
Alien Hunter	83	0	83	17
GIPSy	38	23	15	
GI Hunter	18	12	6	
IslandViewer3	78	41	37	
Zisland Explorer	11	4	7	
Predict Bias	76	25	51	
Sub total	304	105	199	17

FONTE: O autor (2017).
NOTA: Em destaque, a ferramenta com melhor desempenho. [1] (número total de predições entre as ferramentas), [2] (número total de regiões similares que Alien Hunter identificou entre as predições das outras ferramentas), [3] (total das predições entre as ferramentas diminuídas das regiões similares), [4] (regiões únicas que somente a ferramenta identificou).

Alien Hunter conseguiu uma boa cobertura de similaridade, seguido de IslandViewer3. GIPSy obteve número total de suas predições similares próximos de Predict Bias, entretanto, este preditor obteve a maior taxa de ilhas não identificadas por nenhuma outra ferramenta. Zisland Explorer e GI Hunter não possuem nenhuma ilha independente e mostraram número de predições similares baixos quando comparados aos outros preditores. No gráfico a seguir (gráfico 5), esses dados podem ser visualizados.

GRÁFICO 5 - NÚMERO DE GIs PREDITAS, SIMILARES, E ÚNICAS POR CADA FERRAMENTA EM *ESCHERICHIA COLI* CFT073



FONTE: O Autor (2017).

4.6.2 *Streptococcus pneumoniae* R6

O tamanho dos genomas em organismos gram positivos é relativamente menor do que o gram positivo em nosso conjunto de dados, dessa maneira as características das ilhas nestes organismos podem se apresentar de formas diferentes. As ilhas identificadas, apresentam em média, um tamanho de 15 kb, quase o tamanho de uma ilhota genômica. A maior região predita contém 62 genes dentro de 75 kb, e a menor, 3 genes em 4 kb. A soma de todos os resultados das ferramentas é de 162, sendo 86

designadas como ilhas genômicas, e 76 como ilhotas. Na tabela abaixo (tabela 12), é possível visualizar algumas dessas características em cada ferramenta.

TABELA 12 - INFORMAÇÕES DAS ILHAS GENOMICAS EM *STREPTOCOCCUS PNEUMONIAE* R6

Preditores	Média[1]	Mínima[2]	<Genes[3]	Máxima[4]	>Genes[5]	Total Pred[6]	<10 kb[7]	>10 kb[8]
GI Hunter	26296	5729	7	75089	62	15	3	12
Predict Bias	24496	5826	6	54661	51	27	6	21
GIPSy	15118	8167	8	28300	21	12	4	8
Alien Hunter	10588	5148	4	28550	21	79	43	36
Zisland Explorer	11876	4061	3	17656	12	6	2	4
IslandViewer3	6736	4011	3	13581	12	23	18	5

FONTE: O Autor (2017).
NOTA: [1] (Tamanho médio das ilhas preditas por cada ferramenta), [2] (Tamanho mínimo das ilhas preditas por cada ferramenta), [3] (Quantidade de genes presentes nas ilhas de tamanho mínimo por cada ferramenta), [4] (Tamanho máximo das ilhas preditas por cada ferramenta), [5] (Quantidade de genes presentes nas ilhas de tamanho máximo por cada ferramenta), [6] (Total de ilhas preditas por cada ferramenta), [7] (Quantidade total de ilhotas preditas por cada ferramenta), [8] (Quantidade total de GIs preditas por cada ferramenta.

A tabela 13 demonstra os dados para elaboração do gráfico com os resultados similares e regiões únicas identificados pelas ferramentas.

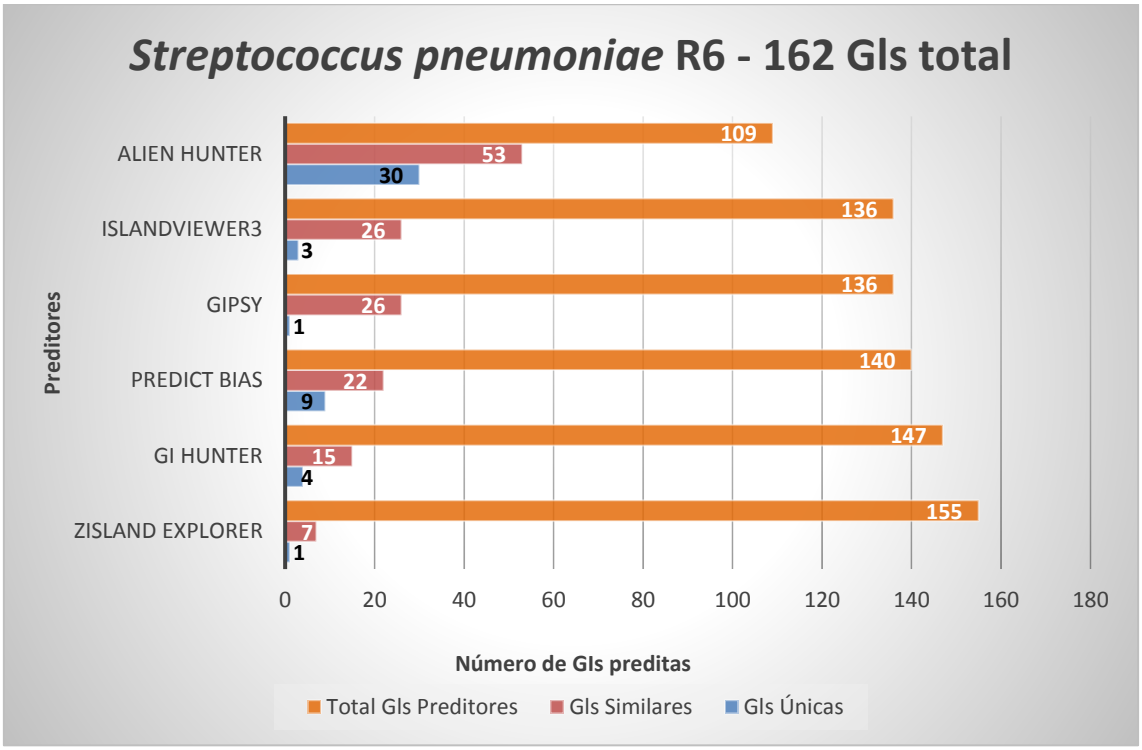
TABELA 13 - PREDITOR COM MELHOR DESEMPENHO EM *STREPTOCOCCUS PNEUMONIAE* R6

Preditores	Total predição[1]	Similares[2]	T. predição s/ similares[3]	GIs Únicas[4]
Alien Hunter	79	0	79	30
GIPSy	12	11	1	
GI Hunter	15	9	6	
IslandViewer3	23	10	13	
Zisland Explorer	6	1	5	
Predict Bias	27	22	5	
Sub total	162	53	109	30

FONTE: O autor (2017).
NOTA: Em destaque, a ferramenta com melhor desempenho. [1] (número total de predições entre as ferramentas), [2] (número total de regiões similares que Alien Hunter identificou entre as predições das outras ferramentas), [3] (total das predições entre as ferramentas diminuídas das regiões similares), [4] (regiões únicas que somente a ferramenta identificou).

Alien Hunter obteve maior número de regiões similares e também de ilhas que somente o mesmo conseguiu reconhecer. IslandViewer3 e GIPSy empataram na detecção de ilhas similares, seguidos de Predict Bias, que ainda se diferencia por ter encontrado mais ilhas únicas do que os preditores citados anteriormente. GI Hunter e Zisland Explorer não tiveram bons resultados comparados com as outras ferramentas. O gráfico a seguir (gráfico 6), demonstra estes resultados.

GRÁFICO 6 - NÚMERO DE GIS PREDITAS, SIMILARES, E ÚNICAS POR CADA FERRAMENTA EM *STREPTOCOCCUS PNEUMONIAE* R6



FONTE: O autor (2017).

4.6.3 *Aeromonas hydrophila* ATCC 7966

A relação média do tamanho das ilhas identificadas neste organismo, possui cerca de 19 kb. A menor ilha contém apenas 2 genes em uma região de 3 kb, e a maior com 97 genes, sobre uma composição de 97 kb.O total de ilhas preditas pelas ferramentas é 157, das quais, 93 são designadas como ilhas genômicas, e 64 como ilhotas, de acordo com o seu tamanho. A tabela a seguir (tabela 14), exemplifica esses resultados.

TABELA 14 - INFORMAÇÕES DAS ILHAS GENÔMICAS EM *AEROMONAS HYDROPHILA* ATCC 7966

Preditores	Média[1]	Mínima[2]	<Genes[3]	Máxima[4]	>Genes[5]	Total Pred[6]	<10 kb[7]	>10 kb[8]
Predict Bias	23563	4136	6	97170	97	78	13	65
IslandViewer3	9133	4000	4	48763	47	13	12	1
GI Hunter	28265	7943	4	47774	42	6	2	4
Zisland Explorer	25501	7344	6	40097	35	8	1	7
Alien Hunter	10302	3220	2	38183	34	40	30	10
GIPSy	18558	5059	6	42592	33	12	6	6

FONTE: O Autor (2017).
NOTA: [1] (Tamanho médio das ilhas preditas por cada ferramenta), [2] (Tamanho mínimo das ilhas preditas por cada ferramenta), [3] (Quantidade de genes presentes nas ilhas de tamanho mínimo por cada ferramenta), [4] (Tamanho máximo das ilhas preditas por cada ferramenta), [5] (Quantidade de genes presentes nas ilhas de tamanho máximo por cada ferramenta), [6] (Total de ilhas preditas por cada ferramenta), [7] (Quantidade total de ilhotas preditas por cada ferramenta), [8] (Quantidade total de GIs preditas por cada ferramenta).

A tabela 15 demonstra os dados para elaboração do gráfico com os resultados similares e regiões únicas identificados pelas ferramentas.

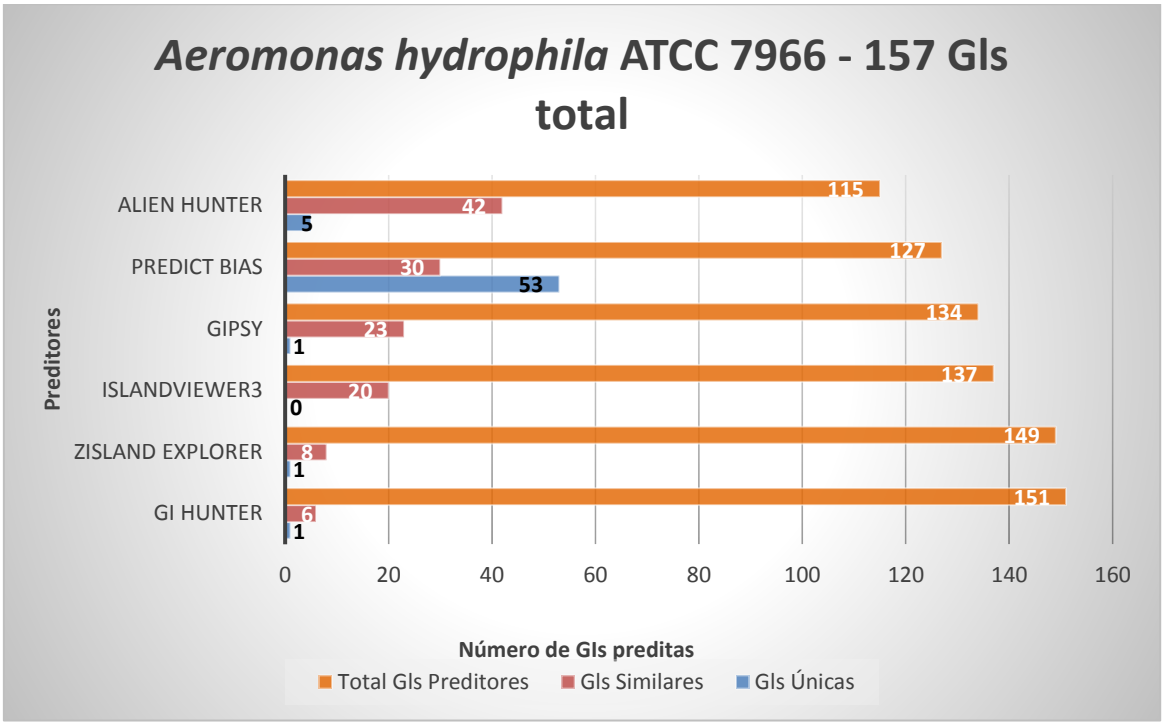
TABELA 15 - PREDITOR COM MELHOR DESEMPENHO EM *AEROMONAS HYDROPHILA* ATCC 7966

Preditores	Total predição[1]	Similares[2]	T. predição s/ similares[3]	GIs Únicas[4]
Alien Hunter	40	0	40	5
GIPSy	12	11	1	
GI Hunter	6	0	6	
IslandViewer3	13	5	8	
Zisland Explorer	8	1	7	
Predict Bias	78	25	53	
Sub total	157	42	115	5

FONTE: O autor (2017).
NOTA: Em destaque a ferramenta com melhor desempenho. [1] (número total de predições entre as ferramentas), [2] (número total de regiões similares que Alien Hunter identificou entre as predições das outras ferramentas), [3] (total das predições entre as ferramentas diminuídas das regiões similares), [4] (regiões únicas que somente a ferramenta identificou).

Os resultados voltados para as ilhas similares da *AEROMONAS HYDROPHILA* ATCC 7966, são os mais parecidos quando comparados entre as ferramentas. Alien Hunter e Predict Bias, identificam quase o mesmo número de ilhas similares entre todos os resultados, entretanto, Predict Bias consegue se destacar por ter predito 53 ilhas únicas, que nenhum outro preditor conseguiu. GIPSy e IslandViewer3, possuem resultados muito semelhantes para ilhas similares, mas somente GIPSy identificou ilha única, porém, é somente uma região. Seguido de Zisland Explorer e GI Hunter, novamente com baixo desempenho entre as ilhas identificadas. Esses dados podem ser visualizados no gráfico a seguir (gráfico 7).

GRÁFICO 7 - NÚMERO DE GIS PREDITAS, SIMILARES E ÚNICAS, POR CADA FERRAMENTA EM *AEROMONAS HYDROPHILA* ATCC 7966



FONTE: O Autor (2017).

5 DISCUSSÃO

Diferentes artigos de revisão buscaram avaliar as ferramentas de predição de ilhas genômicas e suas metodologias empregadas, comparando uma com as outras a partir dos métodos originais de cada uma (LANGILLE et al., 2010; CHE; HASAN; CHEN, 2014; LU; LEONG, 2016; SOARES et al., 2016). Estes artigos de revisão consideraram várias características para seus resultados, como, acurácia, sensibilidade e especificidade das ferramentas, quando confrontados, mas nenhum demonstrou esses resultados utilizando conjuntos de organismos.

Um dos pontos mais importantes que devem ser considerados para um bom resultado durante as predições, é a qualidade das sequências genômicas utilizadas. Com o avanço da tecnologia de sequenciamento, o número de organismos com seus genomas completamente sequenciados cresceu consideravelmente, assim como os *drafts* genomas. Os *drafts* podem ser usados para analisar possíveis GIs candidatas, possibilitando obter informações previamente, porém, vários resultados podem conter falsos positivos e falsos negativos devido ao grande número de *gaps* ao longo das sequências (SOARES et al., 2016).

A única ferramenta presente neste trabalho que possui esta característica é IslandViewer3. Entretanto, os próprios desenvolvedores deixam claro que ainda é preferível o uso de genomas completos para realizar as predições, visando minimizar as possíveis falhas nos resultados. Em nossas análises utilizamos somente genomas completos e suas devidas anotações, provenientes dos arquivos .GBK, e também outros arquivos contendo informações separadas. Como por exemplo, arquivos .PTT, contendo lista das posições de todas as proteínas presentes no genoma, e extensões em .RNT, que possuem dados sobre tRNA e RNAs ribossomais, necessários em algumas ferramentas.

Por mais que as informações contidas nesses arquivos já tenham passado por um devido tratamento nas anotações, em nossas análises foi possível perceber que, algumas CDS presentes em regiões de interesse apresentavam seus produtos como proteínas hipotéticas. Essas proteínas hipotéticas possuem grande importância em PAIs, devido grande parte das CDS presentes nessas ilhas possuírem essa característica, e podem conter funções que ainda não foram determinadas, de acordo com (CHE; HASAN; CHEN, 2014).

Realizando BLAST no UNIPROT, entre algumas dessas CDS de interesse, elas já continham especificações, contribuindo então, para um melhor entendimento dos processos biológicos envolvidos nesta região. Acreditamos que, um dos possíveis motivos para essa falta de informações em alguns produtos, seja devido a data do envio do genoma para depósito. Muitos genes vêm sendo curados manualmente e depositados nos principais bancos de dados no decorrer dos anos. Notamos nas ferramentas analisadas que, nenhuma passa por um pré-processamento dos dados de entrada, como por exemplo, re-anotações dos produtos do genoma nos bancos de dados mais relevantes. É claro que, para isso ocorrer, a conexão com a internet deverá ser necessária e o tempo de análise poderá aumentar, retirando a característica da ferramenta de trabalhar independentemente.

A geração de outros arquivos provenientes do .GBK, por exemplo, também pode influenciar nos resultados finais das predições, como ocorre em IslandViewer3, GIPSy e Zisland Explorer.

Alien Hunter possui certa vantagem neste ponto. Faz uso de somente arquivo .FASTA para sua predição, podendo analisar genomas recém sequenciados, não necessitando de anotações prévias. Predict Bias utiliza somente .GBK, sem a necessidade de qualquer transformação, não alterando os dados proveniente da anotação original. Em avaliações de dados de referência, quando já está determinado os produtos e regiões de interesse, estas perdas de informações podem ser até mesmo ausentes, possibilitando então, uma melhor compreensão dos resultados provenientes das ferramentas.

Em nossas análises das dezesseis ilhas do padrão ouro curadas *in vitro*, todas as regiões foram preditas. Algumas ilhas por mais de um preditor e outras por somente uma ferramenta. GIPSy obteve o melhor resultado, cobrindo cerca de 91% de toda a composição das ilhas de referência, contra 16% da ferramenta que obteve menor desempenho, Zisland Explorer. Alien Hunter também obteve um bom desempenho, em torno de 81% de acertos, seguido de IslandViewer3, com 78%. Outros preditores como Predict Bias tiveram 31% de cobertura e GI Hunter com 17%.

A principal diferença dos resultados entre Alien Hunter e IslandViewer3 em relação a GIPSy, é a ausência dos genes de tRNAs em suas predições, que se encontram associados as ilhas do padrão ouro.

GIPSy faz uso da implementação do software HMMER3 para busca no banco de dados tRNAdb – *Transfer RNA Database*, neste banco consta informações como,

sequências de tRNA e genes de tRNA. Por outro lado, Alien Hunter e IslandViewer3 não possuem nenhuma ferramenta integrada específica para detecção desses genes, podendo ser um dos motivos para a ausência dos mesmos em seus resultados. A ausência de genes tRNAs nas predições de Alien Hunter e IslandViewer3 também podem levar a baixa presença de integrase em seus resultados, devido a maioria desses genes estarem em posições anteriores ou posteriores dos tRNAs. O mesmo segue para os outros três preditores.

Diferente da presença de transposases, as ferramentas IslandViewer3 e Alien Hunter, tiveram resultados mais semelhantes com as da GIPSy, diferenciado somente em 5 genes não encontrados por IslandViewer3, e 10 por Alien Hunter. Para a identificação das transposases, GIPSy faz uso novamente do HMMER3, mas agora sua busca é realizada no banco de dados PFAM. IslandViewer3 faz uso somente da anotação proveniente do próprio genoma e Alien Hunter é independente. Os outros três preditores, novamente tiveram baixo desempenho.

Outro produto importante contido nestas ilhas de referência são as proteínas hipotéticas, devido à grande presença nas PAIs (CHE; HASAN; CHEN, 2014). Na nona ilha, designada como uma PAI, o gene que traz esta característica de patogenicidade se encontra no meio de duas proteínas hipotéticas. Estes produtos podem influenciar nos processos biológicos daquela região, atuando com o gene patogênico. Estas funções em conjunto ainda não podem ser compreendidas por completo devido à falta de anotações dessas proteínas. GIPSy, Alien Hunter e IslandViewer3 tiveram cobertura de identificação superior a 75%, enquanto os outros três preditores não apresentaram bons resultados.

Das dezesseis ilhas do padrão ouro, 8 são caracterizadas como PAIs, pela presença de genes relacionados a fatores de virulência. No total, 25 genes estão associados a esta característica. GIPSy identificou 23 genes, já Alien Hunter e IslandViewer3 conseguiram 24. O gene *fyaC*, se encontra na região que apenas Alien Hunter encontrou (nona ilha), e o gene *tcpC* (décima ilha) unicamente por IslandViewer3. Outros dois preditores conseguiram resultados razoáveis. Predict Bias identificou 19 genes e GI Hunter 15. Somente Zisland Explorer obteve resultado insatisfatório, de um total de 6 genes.

No geral, as ferramentas conseguiram identificar as regiões de interesse, algumas com seu começo de predição ou término em posições que levaram a perdas de produtos, como aconteceu com Zisland Explorer, Precit Bias e GI Hunter. Alien

Hunter e IslandViewer3 proporcionaram resultados satisfatórios, tendo destaque pelas duas ilhas preditas (nona e décima) independentemente do restante das ferramentas, porém, os genes de tRNAs as vezes se encontravam ausentes em seus resultados, juntamente com as integrases. A presença dos genes tRNA e integrases é uma forte característica que a região pode ser uma candidata a ilha genômica (JUHAS et al., 2009)).

GIPSy se destacou por conseguir cobrir quase todas as CDS presentes nas ilhas padrão ouro, mostrando grande desempenho no primeiro momento. Esta ferramenta faz uso de várias metodologias integradas, podendo proporcionar vantagens em relação aos outros preditores, principalmente pela sua característica de classificar as ilhas encontradas em seus resultados de acordo com suas possíveis funções biológicas.

Para comparação dos preditores contra preditores, sem possuir ilhas de referência, realizamos somente a identificação de regiões em torno de 50% de similaridade entre os resultados, em três organismos selecionados. O conteúdo genético dentro das regiões preditas, como número total de CDS, presença de tRNA, integrases, transposases, entre outros, não foram avaliados neste momento.

Além das ilhas de referência do padrão ouro, buscamos identificar todas as possíveis GIs encontradas em *Escherichia coli* CFT073. Alien Hunter conseguiu melhores resultados. No total de 199 predições feitas pelas ferramentas, 105 são similares as regiões que Alien Hunter também identificou, cobrindo cerca de 52% de todas as ilhas identificadas pelas ferramentas, se destacando também por 17 regiões canditadas a GIs que somente ele assimilou.

Alien Hunter estabeleceu *Threshold* (corte limiar) de 11.14 para todo o genoma. O *score* médio de todas as ilhas identificadas ficou em torno de 18,77, o mínimo 11,15, e a região que mais se diferenciou do restante do genoma contém um valor de 58,33. A ilha com menor *score*, é uma ilhota genômica, com 6 genes presentes e um conteúdo G+C% de 44,77%. Um dos produtos contidos nessa ilhota é "*AraC family transcriptional regulator*", que regulam genes com funções diversas, desde o metabolismo do carbono até respostas de estresse à virulência (FROTA el al., 2004). Essas ilhotas genômicas, por mais que contenham poucos genes em relação as ilhas genômicas em geral, podem conter produtos importantes (JAYAPAL et al., 2007).

IslandViewer3 também apresentou bom desempenho. Este preditor conseguiu cobrir cerca de 37 ilhas similares das predições de Alien Hunter, e conseguiu predizer

6 regiões únicas. IslandViewer3 possui uma característica distinta. Em sua metodologia, uma das etapas em suas análises é o método IslandPick. Dependendo do genoma escolhido para comparação, resultados diferentes podem ser obtidos. No *upload* do arquivo para análise, esta ferramenta não possui opção de escolha do genoma para comparação, podendo obter resultados diferentes para outros organismos.

GIPSy obteve 22 ilhas similares ao Alien Hunter, e conseguiu identificar 8 regiões únicas. Esta ferramenta também possui uma característica parecida com IslandViewer3, é necessário o uso de organismos filogeneticamente relacionados, contudo, é possível escolher o genoma para comparação no início da predição. De mesmo modo, organismos diferentes, podem levar a resultados diferentes. O genoma selecionado que foi utilizado em conjunto com a *Escherichia coli* CFT073 foi *Escherichia coli* str. K-12 substr. MG1655, devido ao mesmo ser utilizado na publicação da ferramenta.

Predict Bias conseguiu identificar 18 regiões similares a Alien Hunter, sendo melhores que GI Hunter e Zisland Explorer.

Em *Streptococcus pneumoniae* R6, Alien Hunter também conseguiu resultados melhores quando comparados aos outros preditores. De 109 ilhas preditas pelas outras ferramentas, 53 regiões são similares aos seus resultados, tendo uma cobertura de 48% de acertos. Entre os seus próprios resultados, 30 regiões candidatas a GIs foram identificadas somente por esta ferramenta.

O seu *Threshold* (corte limiar) para este organismo foi de 15.69. Possuindo um *score* médio entre todas as ilhas preditas de 23.64. O valor mínimo atribuído para ilha é de 15, 79, e o valor máximo 56,68. Nesta ilha com *score* mais elevado entre todas as ilhas, se encontra presente genes ribossomais, podendo ser um dos motivos para o seu *score* mais alto.

A segunda ferramenta que apresentou resultados melhores foi IslandViewer3. Contendo 11 ilhas similares aos resultados de Alien Hunter, e três regiões únicas, seguido de GIPSy. Os resultados das duas ferramentas só se diferem devido a GIPSy conter 10 ilhas similares as predições de Alien Hunter e conter uma região única. O genoma utilizado para comparação na GIPSy é *Streptococcus pneumoniae* strain:NT_110_58, devido ao seu conteúdo G+C% ter variação baixa ao genoma de estudo, 39.70% contra 39.80%. Tanto em IslandViewer3 e GIPSy é necessário o uso de genomas de referência para análise e comparação. Para não influenciar nos

resultados ou direcionar os mesmos, escolhendo outros organismos para testes, somente este foi reproduzido nos resultados da GIPSy, devido ao próprio *Streptococcus pneumoniae* R6 ser um organismo de referência. O mesmo ocorre com IslandViewer3, pois não sabemos qual genoma foi estipulado para análise em seu *pipeline*.

O preditor Predict Bias identificou apenas 13 regiões similares ao Alien Hunter, sendo melhor que Zisland Explorer e GI Hunter.

Alien Hunter novamente conseguiu obter melhores resultados quando comparados as outras ferramentas no genoma de *Aeromonas hydrophila* ATCC 7966. Com 115 ilhas identificadas pelos outros preditores, 42 são similares aos seus resultados, gerando cerca de 32% cobertura de acertos. Em suas próprias predições, 5 candidatas a GIs são identificadas somente por esta ferramenta.

Neste genoma, Alien Hunter determinou o *Threshold* (corte limiar) em torno de 16.56, sendo o maior quando comparado aos outros dois genomas anteriores, e seu *score* médio para as ilhas chegou a dobrar de valor, sendo 30,06. O valor mínimo atribuído para ilha é de 16,61, e o valor máximo 62,57. Nesta ilha com *score* mais elevado, entre todas as ilhas, a variação de conteúdo G+C% chegou a 18% comparado com o conteúdo do genoma.

Neste organismo, Predict Bias conseguiu um segundo melhor desempenho. Identificou cerca de 15 ilhas similares aos resultados de Alien Hunter e cerca de 53 regiões únicas. Este grande número de regiões independentes, pode ter sido influenciado pelos resultados das predições totais de IslandViewer3 e GIPSy. As duas ferramentas apresentaram números muito baixos em suas próprias predições. Novamente vem ao caso, a relação do uso de genomas de referência, devido a *Aeromonas hydrophila* ATCC 7966 ser o próprio genoma referencial. Para GIPSy, utilizamos o genoma *Aeromonas hydrophila* YL17, devido ao seu número de genes ser parecidos com o organismo utilizado para teste, 4275 contra 4284, e por seu conteúdo G+C% não apresentar grandes alterações 61.60% / 61.50%.

Nos resultados do próprio banco de IslandViewer3, contém 19 regiões identificadas, candidatas a GIs. Em nossos testes, os resultados das análises nos proporcionaram 13 ilhas, sendo que nenhuma delas apresenta identificação pelo método IslandPick, contra 6 ilhas constatadas no banco.

Entretanto, GIPSy conseguiu identificar 10 regiões similares a Alien Hunter, contendo uma região única, e IslandViewer3, 9 ilhas, não possuindo resultados

independentes. Apesar do seu baixo número de predições totais, ainda apresentaram melhores resultados que os outros dois preditores.

6 CONCLUSÃO

As ferramentas GIPSy, Alien Hunter, IslandViewer3 apresentaram melhores resultados nos testes. No padrão ouro, GIPSy obteve um desempenho melhor que as outras ferramentas, conseguindo chegar a 91% de cobertura de todas as CDS presentes nas ilhas curadas *in vitro*. Alien Hunter e IslandViewer3 tiveram desempenho semelhantes. A diferença dos resultados entre essas duas ferramentas, em comparação a GIPSy, se deu pela perca de genes de tRNA, que estavam associados as ilhas do padrão ouro.

Contudo, as suas regiões identificadas, correspondem a grande parte das CDS presentes nas ilhas. Um ponto importante entre essas ferramentas, é a predição de duas regiões únicas, que cada uma identificou. Sendo as duas ilhas caracterizadas como patogênicas, contendo genes de virulência importantes para a compreensão dos fatores de patogenicidade. Predict Bias, GI Hunter e Zisland Explorer, não tiveram sucesso em suas predições, devido à grande perca de genes característicos das ilhas genômicas.

No decorrer dos testes, percebemos que o desempenho da GIPSy e IslandViewer3 decaíram em relação aos resultados anteriores. Isto pode ser ocasionado pelo fato das duas ferramentas necessitarem de genomas de referência para as suas análises. GIPSy possui uma limitação devido a esta questão, impondo ao usuário genomas de referência para comparação. Suas análises ainda podem ser determinadas com outros organismos, entretanto, genomas que não possuem nenhum organismo semelhante, faz com que a ferramenta fique indisponível para uso. IslandViewer3 possui certa vantagem, pois faz uso de mais dois métodos que não precisam de outro genoma de entrada para realizar suas análises. Contudo, seu desempenho mostrou que a procura por GIs em genomas de referências, podem levar a interpretações duvidosas, devido aos resultados retornados de nossas análises mostrarem diferenças na presença do método IslandPick, comparados aos que estão depositados no próprio banco.

Predict Bias obteve resultados melhores em relação aos organismo gram negativo do que gram positivo. Na busca por regiões similares, teve resultados semelhantes ao da GIPSy em *Escherichia coli CFT073*, e resultados superiores nos dois outros organismos. GI Hunter teve um desempenho muito baixo. Esta ferramenta faz a integração do Alien Hunter em suas análises, e ainda sim obteve resultados

abaixo da média. Um dos possíveis motivos dessa perda de desempenho, pode estar relacionado na qualidade dos dados utilizados para a construção do seu método de predição. O conjunto de treinamento utilizado para elaborar a árvore de decisão, pode conter falsos positivos e falsos negativos. Devido ao uso de *machine learnining* para treinar estes conjuntos, ocorrências de *overfitting*, podem leva resultados duvidosos. Zisland Explorer não teve muito impacto na comparação com as outras ferramentas. Tanto a aplicação *web*, quanto a *in house*, trazem certas diferenças em suas predições, quando se faz uso do arquivo .PTT em conjunto com .FNA, em relação ao uso do .FNA independente, ficando difícil de avaliar os seus resultados.

Alien Hunter se saiu melhor nos três organismos. Possui a grande vantagem de poder ser utilizado em genomas recém sequenciados, não necessitando de pré-anotações e genomas de referência para consulta. Outra característica importante é a identificação de regiões que possuem genes ribossomais associados, demonstrados em seus arquivos de saída, diminuindo assim, resultados falsos positivos. Sua capacidade de impor valores para regiões candidatas a GIs, a partir da diferença do restante do genoma, é muito importante para pesquisas naquela região de interesse, quanto para assimilar funções biológicas dos genes que se encontram nas proximidades.

A partir de nossos resultados, observamos que ainda é necessário a utilização de mais de um método de predição para as ilhas genômicas. As combinações de várias metodologias podem proporcionar resultados mais satisfatórios, em relação ao uso de somente uma ferramenta. Uma boa estratégia, baseada nos resultados e características distintas das ferramentas avaliadas, é a utilização em conjunto de três ferramentas.

Em primeiro lugar, com o uso de Alien Hunter, o usuário terá uma visão geral de todo o genoma e suas possíveis regiões de interesse, com valores atribuídos. Segundo, IslandViewer3 proporciona uma pesquisa mais interativa e dinâmica. Os genes presentes nas GIs candidatas podem ser rapidamente visualizados e consultados no NCBI, além de trazer informações sobre possíveis genes associados a fatores de patogenicidade e resistência a antibióticos. Em um terceiro momento, o uso da ferramenta GIPSy. Devido a suas várias metodologias integradas, principalmente pela identificação de tRNAs e sua capacidade de classificar as ilhas de acordo com as suas possíveis funções.

Com essas informações, o usuário poderá obter os dados para efetuar análises mais aprofundadas sobre as regiões de interesse, candidatas a GIs. A interpretação e compreensão das funções relacionadas as ilhas genômicas, podem levar a estudos significativos, relacionados aos principais fatores que possibilitam os organismos a evoluir e se adaptarem constantemente.

REFERÊNCIAS

Babić, A., Lindner, A. B., Vulić, M., Stewart, E. J., & Radman, M. (2008). Direct Visualization of Horizontal Gene Transfer. **Science**, 319(5869), 1533–1536. doi.org:10.1126.

Bellanger, X., Payot, S., Leblond-Bourget, N., & Guédon, G. (2014). Conjugative and mobilizable genomic islands in bacteria: Evolution and diversity. **FEMS Microbiology Reviews**, 38(4), 720–760. https://doi.org/10.1111/1574-6976.12058.

Berg, O. G., & Kurland, C. G. (2002). Evolution of microbial genomes: sequence acquisition and loss. **Molecular Biology and Evolution**, 19(12), 2265–76. doi.org:10.1093.

Buchrieser, C., Prentice, M., & Carniel, E. (1998). The 102-kilobase unstable region of Yersinia pestis comprises a high- pathogenicity island linked to a pigmentation segment which undergoes internal rearrangement. **Journal of Bacteriology**, 180(9), 2321–2329.

Burrus, V., Pavlovic, G., Decaris, B., & Guédon, G. (2002). Conjugative transposons: The tip of the iceberg. **Molecular Microbiology**, 46(3), 601–610. doi.org:10.1046.

Burrus, V., & Waldor, M. K. (2004). Shaping bacterial genomes with integrative and conjugative elements. **Research in Microbiology**, 155(5), 376–386. doi.org:10.1016.
Cambray, G., Guerout, A.-M., & Mazel, D. (2010). Integrons. **Annual Review of Genetics**, 44(1), 141–166. doi.org:10.1146.

Canchaya, C., Proux, C., Fournous, G., Bruttin, A., & Brüssow, H. (2003). Prophage genomics. **Microbiology and Molecular Biology Reviews : MMBR**, 67(2), 238–276, table of contents. doi.org/10.1128.

Carver, T., Harris, S. R., Berriman, M., Parkhill, J., & McQuillan, J. A. (2012). Artemis: An integrated platform for visualization and analysis of high-throughput sequence-based experimental data. **Bioinformatics,** 28(4), 464–469. doi.org:10.1093.

Che, D., Hasan, M. S., & Chen, B. (2014). Identifying pathogenicity islands in bacterial pathogenomics using computational approaches. **Pathogens** *(Basel, Switzerland)*, *3*(1), 36–56. Do.orgi:10.3390.

Che, D., & Wang, H. (2014). GIV: A Tool for Genomic Islands Visualization. **Bioinformation**, 9(17), 879–82. doi.org:10.6026.

Che, D., Wang, H., Fazekas, J. (2014). An Accurate Genomic Island Prediction Method for Sequenced Bacterial and Archaeal Genomes. **Journal of Proteomics & Bioinformatics**, *7*(8), 214–221. doi.org:10.4172.

Chen, I., & Dubnau, D. (2004). DNA uptake during bacterial transformation. **Nature Reviews. Microbiology**, 2(3), 241–249. doi.org:10.1038.

Datsenko, K. A., & Wanner, B. L. (2000). One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. **Proceedings of the National Academy of Sciences of the United States of America**, 97(12), 6640–5. doi.org/10.1073.

Darling, A. C. E., Mau, B., Blattner, F. R., & Perna, N. T. (2004). Mauve : Multiple Alignment of Conserved Genomic Sequence With Rearrangements. **Genome Research**, 1394–1403. doi.org:10.1101.

Dhillon, B. K., Laird, M. R., Shay, J. A., Winsor, G. L., Lo, R., Nizam, F., Brinkman, F. S. L. (2015). IslandViewer 3: More flexible, interactive genomic island discovery, visualization and analysis. **Nucleic Acids Research**, *43*(W1), W104–W108. doi.org:10.1093

Dobrindt, U., Hochhut, B., Hentschel, U., & Hacker, J. (2004). Genomic islands in pathogenic and environmental microorganisms. **Nature Reviews Microbiology**, 2(5), 414–424. doi.org:10.1038.

Eddy, S. R. (2011). Accelerated profile HMM searches. **PLoS Computational Biology**, *7*(10). doi.org:10.1371.

Erjavec, M. S., Jesenko, B., Petkovšek, Ž., & Žgur-Bertok, D. (2010). Prevalence and associations of tcpC, a gene encoding a Toll/interleukin-1 receptor domain-containing protein, among Escherichia coli urinary tract infection, skin and soft tissue infection, and commensal isolates. **Journal of Clinical Microbiology**, 48(3), 966–968. doi.org:10.1128.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Punta, M. (2013). Pfam: The protein families database. **Nucleic Acids Research**, *42*(D1), 222–230. doi.org:10.1093.

Frank, E., Hall, M. A., & Witten, I. H. (2016). The WEKA Workbench. Online Appendix. Data Mining: Practical Machine Learning Tools and Techniques. **Morgan Kaufmann**, Fourth Edition.

Frota, C. C., Papavinasasundaram, K. G., Davis, E. O., & Colston, M. J. (2004). The AraC family transcriptional regulator Rv1931c plays a role in the virulence of Mycobacterium tuberculosis. **Infection and Immunity**, 72(9), 5483–5486. doi.org:10.1128.

Frost, L. S., Leplae, R., Summers, A. O., & Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. **Nat.Rev.Microbiol**., 3(1740–1526, 722–732. doi.org:10.1038.

Gal-Mor, O., & Finlay, B. B. (2006). Pathogenicity islands: A molecular toolbox for bacterial virulence. **Cellular Microbiology**, 8(11), 1707–1719. doi.org:10.1111.

Gao, F., & Zhang, C. T. (2006). GC-Profile: A web-based tool for visualizing and analyzing the variation of GC content in genomic sequences. **Nucleic Acids Research**, 34(WEB. SERV. ISS.), 686–691. doi.org:10.1093.

Grohmann, E., Muth, G., & Espinosa, M. (2003). Conjugative plasmid transfer in gram-positive bacteria. **Microbiology and Molecular Biology Reviews** : MMBR, 67(2), 277–301. doi.org:10.1128.

Groisman, E. A., & Ochman, H. (1996). Pathogenicity islands: Bacterial evolution in quantum leaps. **Cell**, 87(5), 791–794. doi.org/10.1016.

Griffith, F. (1981). Classics in infectious diseases: The significance of pneumococcal types. **Reviews of Infectious Diseases**, 3(2), 372–395. doi.org:10.1093.

Hacker, J., Bender, L., Ott, M., Wingender, J., Lund, B., Marre, R., & Goebel, W. (1990). Deletions of chromosomal regions coding for fimbriae and hemolysins occur in vitro and in vivo in various extra intestinal Escherichia coli isolates. **Microbial Pathogenesis**, 8(3), 213–225. doi.org:10.1016.

Hacker, J., Blum-Oehler, G., Mühldorfer, I., & Tschäpe, H. (1997). Pathogenicity islands of virulent bacteria: Structure, function and impact on microbial evolution. **Molecular Microbiology**, 23(6), 1089–1097. doi.org:10.1046.

Hacker, J., Carniel, E., Achtmann, M., Zurth, K., Morelli, G., Torrea, G., Mekalanos, J. J. (2001). Ecological fitness, genomic islands and bacterial pathogenicity A Darwinian view of the evolution of microbes. **EMBO Reports**, 2(5), 376–381. doi.org:10.1093.

Hacker J, Bender L, Ott M, Wingender J, Lund B, Marre R, Goebel W. Deletions of chromosomal regions coding for fimbriae and hemolysin occurr in vitro and in vivo in various extraintestinal Escherichia coli. **Microbial Pathog**. 1990;8:213-25

Hancock, V., Ferrières, L., & Klemm, P. (2008). The ferric yersiniabactin uptake receptor FyuA is required for efficient biofilm formation by urinary tract infectious Escherichia coli in human urine. **Microbiology**, 154(1), 167–175. doi.org:10.1099.

Hacker, J., & Kaper, J. B. (2000). Pathogenicity Islands and the Evolution of Microbes. **Annu. Rev. Microbiol**, 54, 641–79. doi.org:10.1146.

Hensel, M. (2004). Evolution of pathogenicity islands of Salmonella enterica. **International Journal of Medical Microbiology**, 294(2–3), 95–102.

Hsiao, W. W. L., Ung, K., Aeschliman, D., Bryan, J., Brett Finlay, B., & Brinkman, F. S. L. (2005). Evidence of a large novel gene pool associated with prokaryotic genomic Islands. **PLoS Genetics**, 1(5), 540–550. doi.org:10.1371.

Ho Sui, S. J., Fedynak, A., Hsiao, W. W. L., Langille, M. G. I., & Brinkman, F. S. L. (2009). The association of virulence factors with genomic Islands. **PLoS ONE**, *4*(12). doi.org:10.1371

Hudson, C. M., Lau, B. Y., & Williams, K. P. (2015). Islander: A database of precisely mapped genomic islands in tRNA and tmRNA genes. **Nucleic Acids Research**, 43(D1), D48–D53. doi.org:10.1093.

Hsiao, W., Wan, I., Jones, S. J., & Brinkman, F. S. L. (2003). IslandPath : aiding detection of genomic islands in prokaryotes, **BMC Bioinformatics** 19(3), 418–420. doi.org:10.1093.

Jayapal, K. P., Lian, W., Glod, F., Sherman, D. H., & Hu, W.-S. (2007). Comparative genomic hybridizations reveal absence of large Streptomyces coelicolor genomic islands in Streptomyces lividans. **BMC Genomics**, 8, 229. doi.org:10.1186.

Juhas, M., Crook, D. W., Dimopoulou, I. D., Lunter, G., Harding, R. M., Ferguson, D.

J. P., & Hood, D. W. (2007). Novel type IV secretion system involved in propagation of genomic islands. **Journal of Bacteriology**, 189(3), 761–771. doi.org:10.1128.

Juhas, M., Van Der Meer, J. R., Gaillard, M., Harding, R. M., Hood, D. W., & Crook, D. W. (2009). Genomic islands: Tools of bacterial horizontal gene transfer and evolution. **FEMS Microbiology Reviews**, 33(2), 376–393. doi: 10.1111.

Jühling, F., Mörl, M., Hartmann, R. K., Sprinzl, M., Stadler, P. F., & Pütz, J. (2009). tRNAdb 2009: Compilation of tRNA sequences and tRNA genes. **Nucleic Acids Research**, *37*(SUPPL. 1), 159–162. doi.org:10.1093.

KAPER, J.B.; HACKER, J.; **Pathogenicity Islands and other mobile virulence elements**. ASM Press. Washington, 1999, 366p.

Karlin, S. (2001). Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. **Trends in Microbiology**, 9(7), 335–343. doi.org:10.1016.

Koonin, E. V., Makarova, K. S., & Aravind, L. (2001). Horizontal gene transfer in prokaryotes: quantification and classification. **Annual Review of Microbiology**, 55(709), 709–42. doi.org:10.1146.

Langille, M. G., Hsiao, W. W., & Brinkman, F. S. (2008). Evaluation of genomic island predictors using a comparative genomics approach. **BMC Bioinformatics**, 9, 329. doi.org:10.1186

Langille, M. G., Hsiao, W. W., & Brinkman, F. S. (2010). Detecting genomic islands using bioinformatics approaches. **Nat Rev Microbiol**, *8*(5), 373–382. doi: 10.1038.

Lee, C. C., Chen, Y. P. P., Yao, T. J., Ma, C. Y., Lo, W. C., Lyu, P. C., & Tang, C. Y. (2013). GI-POP: A combinational annotation and genomic island prediction pipeline for ongoing microbial genome projects. **Gene**, 518(1), 114–123. doi.org:10.1016.

Lederberg, J. and Tatum, E.L. (1946) Novel genotypes in mixed cultures of biochemical mutants of bacteia, **Cold Spring Harbor Symposia on Quantitative Biology**, 11, 113-114.

Liu, B., & Pop, M. (2009). ARDB--Antibiotic Resistance Genes Database. **Nucleic Acids Research**, *37*(Database issue), D443–D447. doi:10.1093.

Lloyd, A. L., Rasko, D. A., & Mobley, H. L. T. (2007). Defining genomic islands and uropathogen-specific genes in uropathogenic Escherichia coli. **Journal of Bacteriology**, 189(9), 3532–3546. doi.org:10.1128.

Lloyd, A. L., Henderson, T. A., Vigil, P. D., & Mobley, H. L. T. (2009). Genomic islands of uropathogenic Escherichia coli contribute to virulence. **Journal of Bacteriology**, 191(11), 3469–3481. doi.org:10.1128.

Lorenz, M. G., & Wackernagel, W. (1994). Bacterial gene transfer by natural genetic transformation in the environment. **Microbiological Reviews**, 58(3), 563–602.

Lu, B., & Leong, H. W. (2016). Computational methods for predicting genomic islands in microbial genomes. **Computational and Structural Biotechnology Journal**, *14*, 200–206. doi:10.1016.

Lwoff, A. (1953). Lysogeny. **Bacteriological Reviews**, 17(4), 269–337.

Maiden, M. C. J. (1998). Horizontal Genetic Exchange, Evolution, and Spread of Antibiotic Resistance in Bacteria, *27*(Suppl 1), 12–20.

Mantri, Y., & Williams, K. P. (2004). Islander: a database of integrative islands in prokaryotic genomes, the associated integrases and their DNA site specificities. **Nucleic Acids Res**, 32(Database issue), D55-8. doi.org:10.1093.

Mao, C., Qiu, J., Wang, C., Charles, T. C., & Sobral, B. W. S. (2005). NodMutDB: A database for genes and mutants involved in symbiosis. **Bioinformatics**, *21*(12), 2927–2929. doi.org:10.1093.

Maurice, C. F., Bouvier, C., De Wit, R., & Bouvier, T. (2013). Linking the lytic and lysogenic bacteriophage cycles to environmental conditions, host physiology and their variability in coastal lagoons. **Environmental Microbiology,** 15(9), 2463–2475. doi.org:10.1111.

McArthur, A. G., Waglechner, N., Nizam, F., Yan, A., Azad, M. A., Baylay, A. J., Wright, G. D. (2013). The comprehensive antibiotic resistance database. **Antimicrobial Agents and Chemotherapy**, *57*(7), 3348–3357. doi.org:10.1128

Nielsen, K. M., Bøhn, T., & Townsend, J. P. (2013). Detecting rare gene transfer events in bacterial populations. **Frontiers in Microbiology**, 4(JAN), 1–12. doi: 10.3389.

Novick, R. P., & York, N. (2013). Pathogenicity and Other Genomic Islands. **Brenner's Encyclopedia of Genetics, Second Edition** (Vol. 5). Elsevier Inc. doi.org:10.1016.

Pierneef, R., Cronje, L., Bezuidt, O., & Reva, O. N. (2015). Pre _ GI : a global map of ontological links between horizontally transferred genomic islands in bacterial and archaeal genomes. **Databses**, (3), 1–13. doi.org/10.1093.

Pundhir, S., Vijayvargiya, H., & Kumar, A. (2008). PredictBias: a server for the identification of genomic and pathogenicity islands in prokaryotes. **In Silico Biology**, 8(April), 223–234. doi.org:2008080019.

Rankin, D. J., Rocha, E. P. C., & Brown, S. P. (2010). What traits are carried on mobile genetic elements , and why ? **Heredity**, 106(1), 1–10. doi.org/10.1038.

Rissman, A. I., Mau, B., Biehl, B. S., Darling, A. E., Glasner, J. D., & Perna, N. T. (2009). Reordering contigs of draft genomes using the Mauve Aligner. **Bioinformatics**, 25(16), 2071–2073. doi.org/10.1093.

Sampaio, RF, & Mancini, MC. (2007). Estudos de revisão sistemática: um guia para síntese criteriosa da evidência científica. ***Brazilian Journal of Physical Therapy*****,** *11*(1), 83-89. https://dx.doi.org/10.1590

Schmidt, H., & Hensel, M. (2004). Pathogenicity islands in bacterial pathogenesis. **Clin.Microbiol.Rev**., 17(0893–8512, 14–56. doi.org:10.1128.

Siguier, P., Perochon, J., Lestrade, L., Mahillon, J., & Chandler, M. (2006). ISfinder : the reference centre for bacterial insertion sequences, **Nucleic Acids Res** 34, 32–36. doi.org:10.1093.

Soares, S. C., AbreuViní, V. A. C., Ramos, R. T. J., Cerdeira, L., Silva, A., Baumbach, J., Azevedo, V. (2012). PIPS: Pathogenicity island prediction software. **PLoS ONE**, *7*(2). doi.org:10.1371.

Soares, S. C., Geyik, H., Ramos, R. T. J., de Sá, P. H. C. G., Barbosa, E. G. V, Baumbach, J., Azevedo, V. (2016). GIPSy: Genomic island prediction software. **Journal of Biotechnology**, 232, 2–11. doi.org: 10.1016.

Soares, G. M. S., Figueiredo, L. C., Faveri, M., Cortelli, S. C., Duarte, P. M., & Feres, M. (2012). Mechanisms of action of systemic antibiotics used in periodontal treatment and mechanisms of bacterial resistance to these drugs. **Journal of Applied Oral Science**, 20(3), 295–309. doi: 10.1590.

Soares, S. D. C., Oliveira, L. D. C., & Jaiswal, A. K. (2016). Genomic Islands : an overview of current software tools and future improvements, **Journal of Integrative Bioinformatics**, 13(1). doi.org/10.2390.

Tatum, E. L., & Lederberg, J. (1947). Gene Recombination in the Bacterium Escherichia coli. **Journal of Bacteriology**, 53(6), 673–684. doi.org:10.1038.

Thomas, C. M. (2000). Paradigms of plasmid organization. **Molecular Microbiology**, 37(3), 485–491. doi.org:10.1046.

Vejborg RM, Hancock V, Schembri MA, Klemm P. Comparative genomics of Escherichia coli strains causing urinary tract infections. **Appl Environ Microbiol**. 2011;77:3268–3278. doi: 10.1128.

Vernikos, G. S., & Parkhill, J. (2006). Interpolated variable order motifs for identification of horizontally acquired DNA: Revisiting the Salmonella pathogenicity islands. **Bioinformatics**, *22*(18), 2196–2203. doi:10.1093.

Von Wintersdorff, C. J. H., Penders, J., Van Niekerk, J. M., Mills, N. D., Majumder, S., Van Alphen, L. B., Wolffs, P. F. G. (2016). Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer. **Frontiers in Microbiology**, *7*(FEB), 1–10. doi.org:10.3389.

Waack, S., Keller, O., Asper, R., Brodag, T., Damm, C., Fricke, W. F., Merkl, R. (2006). Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. **BMC Bioinformatics**, *7*, 142. doi.org:10.1186.

Wattam, A. R., Abraham, D., Dalay, O., Disz, T. L., Driscoll, T., Gabbard, J. L., Sobral, B. W. (2014). PATRIC, the bacterial bioinformatics database and analysis resource. **Nucleic Acids Research**, *42*(D1), 581–591. doi.org:10.1093.

Wei, W., Gao, F., Du, M.-Z., Hua, H.-L., Wang, J., & Guo, F.-B. (2016). Zisland Explorer: detect genomic islands by combining homogeneity and heterogeneity properties. **Briefings in Bioinformatics**, (January), bbw019. doi.org:10.1093.

Wilson, D. J. (2012). Insights from Genomics into Bacterial Pathogen Populations. **PLoS Pathogens**, 8(9). doi: 10.1371.

Xu, Z., & Hao, B. (2009). CVTree update: A newly designed phylogenetic study platform using composition vectors and whole genomes. **Nucleic Acids Research**, 37(SUPPL. 2), 174–178. doi.org:10.1093.

Yoon, S. H., Park, Y. K., & Kim, J. F. (2015). PAIDB v2.0: Exploration and analysis of pathogenicity and resistance islands. **Nucleic Acids Research**, 43(D1), D624–D630. doi.org/10.1093.

Zhou, C. E., Smith, J., Lam, M., Zemla, A., Dyer, M. D., & Slezak, T. (2007). MvirDB - A microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. **Nucleic Acids Research**, *35*(SUPPL. 1), 391–394. doi.org:10.1093.

ANEXO 1 DESCRIÇÃO TAXONÔMICA DOS ORGANISMOS SELECIONADOS

Filo	Classe	Ordem	Família	Gênero	Espécie	Estirpe	Gram +/-
Actinobacteria	Actinobacteria	Corynebacteriales	Corynebacteriaceae	Corynebacterium	*Corynebacterium diphtheriae*	NCTC 13129	+
Actinobacteria	Actinobacteria	Corynebacteriales	Corynebacteriaceae	Corynebacterium	*Corynebacterium glutamicum*	ATCC 13032	+
Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus	*Streptococcus agalactiae*	NEM316	+
Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus	*Streptococcus mitis*	B6	+
Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus	*Streptococcus pyogenes*	M1 GAS	+
Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus	*Streptococcus pneumoniae*	R6	+
Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Staphylococcus	*Staphylococcus aureus*	NCTC 8325	+
Proteobacteria	Gammaproteobacteria	Enterobacterales	Enterobacteriaceae	Escherichia	*Escherichia coli*	K-12 substr. MG1655	-
Proteobacteria	Gammaproteobacteria	Enterobacterales	Enterobacteriaceae	Escherichia	*Escherichia coli*	CFT073	-
Proteobacteria	Gammaproteobacteria	Enterobacterales	Enterobacteriaceae	Escherichia	*Escherichia coli*	Sakai	-
Proteobacteria	Gammaproteobacteria	Enterobacterales	Enterobacteriaceae	Klebsiella	*Klebsiella pneumoniae*	HS11286	-
Proteobacteria	Gammaproteobacteria	Enterobacterales	Enterobacteriaceae	Salmonella	*Salmonella enterica*	CT18	-
Proteobacteria	Gammaproteobacteria	Aeromonadales	Aeromonadaceae	Aeromonas	*Aeromonas hydrophila*	ATCC 7966	-
Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas	*Pseudomonas aeruginosa*	PAO1	-
Proteobacteria	Gammaproteobacteria	Vibrionales	Vibrionaceae	Vibrio	*Vibrio cholerae*	N16961	-
Proteobacteria	Gammaproteobacteria	Vibrionales	Vibrionaceae	Vibrio	*Vibrio cholerae*	N16961	-

ANEXO 2 DESCRIÇÃO COMPLEMENTAR DOS DADOS DOS ORGANISMOS E EXTENSÕES BAIXADAS

Organismo	N. de acesso NCBI	Envio	Data Submissão	Data Download
Corynebacterium diphtheriae NCTC 13129	NC_002935.2	Sanger Institute	10/11/03	07/06/16
Corynebacterium glutamicum ATCC 13032	NC_003450.3	Kitasato Univ.	18/02/04	07/06/16
Streptococcus agalactiae NEM316	NC_004368.1	Institut Pasteur	06/05/03	09/01/17
Streptococcus mitis B6	NC_013853.1	Univ. Kaiserslautern	16/02/10	09/01/17
Streptococcus pyogenes M1 GAS	NC_002737.2	Univ. Oklahoma	01/04/14	13/12/16
Streptococcus pneumoniae R6	NC_003098.1	Eli Lilly	25/11/02	13/12/16
Staphylococcus aureus subsp. aureus NCTC 8325	NC_007795.1	Univ. Oklahoma	13/02/06	13/12/16
Escherichia coli str. K-12 substr. MG1655	NC_000913.3	Univ. Wisconsin	26/09/13	07/06/16
Escherichia coli CFT073	NC_004431.1	Univ. Wisconsin	06/12/02	07/06/16
Escherichia coli O157:H7 Sakai	NC_002695.1	GIRC	11/05/04	07/06/16
Klebsiella pneumoniae subsp. pneumoniae HS11286	NC_016845.1	Univ. Shangai	27/12/11	13/12/16
Salmonella enterica subsp. enterica serovar Typhi str. CT18	NC_003198.1	Sanger Institute	06/05/03	07/06/16
Aeromonas hydrophila subsp. hydrophila ATCC 7966	NC_008570.1	TIGR	06/11/06	13/12/16
Pseudomonas aeruginosa PAO1	NC_002516.2	Genesis Corporation	07/07/06	07/06/16
Vibrio cholerae O1 biovar El Tor N16961 chromosome I	NC_002505.1	TIGR	09/01/01	07/06/16
Vibrio cholerae O1 biovar El Tor N16961 chromosome II	NC_002506.1	TIGR	09/01/01	07/06/16

NOTA: Todos os genomas relacionados apresentam montagem completa. Para todos os organismos relacionados, foram baixados do servidor de arquivos do NCBI os arquivos com as extensões: GBK/.FASTA/.FNA/.PTT/.RNT

ANEXO 3 DESCRIÇÃO QUALITATIVA DOS SOFTWARES PREDITORES DE ILHAS GENÔMICAS ESCOLHIDOS

Preditor	Plataforma	Extensão dos arquivos de entrada	Extensão dos arquivos de saída	Visualização Gráficos	Interpretação dos Resultados	Prediz Função das GIs	Utilização
Alien Hunter	Linux	FASTA	TXT / PLOS / SCO	Compatível com Artemis	Posição Inicial e Final da Ilha	Não	Linha de Comando
GI Hunter	Linux	FNA/PTT/RNT	TXT / PLOT	Compatível com GIV	Posição Inicial e Final da Ilha	Não	Linha de Comando
GIPSy	Linux / Windows	GBK/EMBL	TXT	Não possui	Posição, produtos, classes	Sim (PAI, RIs, MIs, SIs)	GUI
IslandViewer 3	Web	GBK/EMBL	GBK / FASTA / PLOT	Gráfico Interativo Próprio	Posição, produtos, genes associados	Não	GUI
Predict Bias	Web	GBK	TXT / PLOT / HTML	Gráfico Próprio desvio de G+C%	Posição locus tag, produtos, classe PAI	Sim (Somente Patogênica)	GUI
Zisland Explorer	Linux / Windows / Web	FNA/PTT	TXT / PLOT	Gráfico Próprio desvio de G+C%	Posição Inicial e Final da Ilha	Não	GUI

ANEXO 4 RESULTADO TOTAL DE GIS PREDITAS ENTRE OS ORGANISMOS DE CADA FERRAMENTA

Organismo	Alien Hunter	Predict Bias	IslandViewer 3	GI Hunter	Zisland	GIPSy
Corynebacterium diphtheriae NCTC 13129	84	22	35	14	6	28
Corynebacterium glutamicum ATCC 13032	52	33	28	15	2	48
Streptococcus agalactiae NEM316	34	19	12	8	2	17
Streptococcus mitis B6	66	34	24	15	1	23
Streptococcus pyogenes M1 GAS	48	26	4	10	1	5
Streptococcus pneumoniae R6	79	27	23	15	6	12
Staphylococcus aureus subsp. aureus NCTC 8325	10	29	10	8	1	12
Escherichia coli str. K-12 substr. MG1655	61	67	52	18	9	24
Escherichia coli CFT073	83	76	78	18	11	38
Escherichia coli O157:H7 Sakai	96	72	60	23	9	52
Klebsiella pneumoniae subsp. pneumoniae HS11286	49	90	17	18	13	20
Salmonella enterica subsp. enterica Typhi str. CT18	80	59	50	22	6	17
Aeromonas hydrophila subsp. hydrophila ATCC 7966	40	78	13	6	8	12
Pseudomonas aeruginosa PAO1	34	73	23	12	6	7
Vibrio cholerae O1 biovar El Tor N16961 chomo I	24	26	16	8	3	2
Vibrio cholerae O1 biovar El Tor N16961 chromo II	17	12	8	5	1	6
TOTAL	**857**	**743**	**453**	**215**	**85**	**323**

ANEXO 5 RESULTADO TOTAL DE GIS PREDITAS ENTRE OS ORGANISMOS GRAM POSITOVOS E GRAM NEGATIVOS

Organismo	Alien Hunter	Predict Bias	IslandViewer 3	GI Hunter	Zisland Explorer	GIPSy
Grams Positivos						
Corynebacterium diphtheriae NCTC 13129	84	22	35	14	6	28
Corynebacterium glutamicum ATCC 13032	52	33	28	15	2	48
Streptococcus agalactiae NEM316	34	19	12	8	2	17
Streptococcus mitis B6	66	34	24	15	1	23
Streptococcus pyogenes M1 GAS	48	26	4	10	1	5
Streptococcus pneumoniae R6	79	27	23	15	6	12
Staphylococcus aureus subsp. aureus NCTC 8325	10	29	10	8	1	12
TOTAL	373	190	136	85	19	145
Grams Negativos						
Escherichia coli str. K-12 substr. MG1655	61	67	52	18	9	24
Escherichia coli CFT073	83	76	78	18	11	38
Escherichia coli O157:H7 Sakai	96	72	60	23	9	52
Klebsiella pneumoniae subsp. pneumoniae HS11286	49	90	17	18	13	20
Salmonella enterica subsp. enterica serovar Typhi str. CT18	80	59	50	22	6	17
Aeromonas hydrophila subsp. hydrophila ATCC 7966	40	78	13	6	8	12
Pseudomonas aeruginosa PAO1	34	73	23	12	6	7
Vibrio cholerae O1 biovar El Tor N16961 chromosome I	24	26	16	8	3	2
Vibrio cholerae O1 biovar El Tor N16961 chromosome II	17	12	8	5	1	6
TOTAL	**484**	**553**	**317**	**130**	**66**	**178**

ANEXO 6 DESCRIÇÃO DA COMPOSIÇÃO GENOMAS ESCOLHIDOS

Organismo	Tamanho do genoma (pb)	G+C%	Total Genes	Total CDS	rRNA	tRNA	Integrases	Transposases	Genes hipotéticos	Pseudo Gene
Corynebacterium diphtheriae NCTC 13129	2.488.635	53.50%	2,291	2,223	15	52	14	40	553	38
Corynebacterium glutamicum ATCC 13032	3.309.401	53.80%	3,057	2,959	18	60	3	20	1,193	18
Streptococcus agalactiae NEM316	2.211.485	35.60%	2,217	2,107	21	80	8	19	496	13
Streptococcus mitis B6	2.146.611	40.00%	2,096	2,002	12	61	9	62	640	12
Streptococcus pyogenes M1 GAS	1.852.433	38.50%	1,801	1,693	12	60	6 Putativas	8 Putativas	628	0
Streptococcus pneumoniae R6	2.038.615	39.70%	1,967	1,814	12	58	2	19	634	79
Staphylococcus aureus subsp. aureus NCTC 8325	2.821.361	32.90%	2,872	2,767	16	61	8	2	1,496	30
Escherichia coli str. K-12 substr. MG1655	4.641.652	50.80%	4,498	4,329	22	89	7	49	6	184
Escherichia coli CFT073	5.231.428	50.50%	5,179	5,070	21	88	27	108	972	173
Escherichia coli O157:H7 Sakai	5.498.450	50.50%	5,36	5,204	22	105	17	44	1,704	14
Klebsiella pneumoniae subsp. pneumoniae HS11286	5.333.942	57.50%	5,404	5,316	25	62	8	17	1,436	0
Salmonella enterica subsp. enterica serovar Typhi str. CT18	4.809.037	52.10%	4,455	4,111	22	79	10	29	756	233
Aeromonas hydrophila subsp. hydrophila ATCC 7966	4.744.448	61.50%	4,284	4,119	31	128	2	0	856	7
Pseudomonas aeruginosa PAO1	6.264.404	66.60%	5,697	5,572	13	63	1	7	2,256	19
Vibrio cholerae O1 biovar El Tor N16961 chromosome I	2.961.149	47.70%	2,69	2,534	25	94	3	9	716	37