Universidade Federal da Paraíba

Centro de Informática

Programa de Pós-Graduação em Informática

Uma nova abordagem para a identificação de ilhas genômicas em bactérias com base no método de agrupamento *mean shift*

Daniel Miranda de Brito

João Pessoa - PB

Fevereiro de 2017

Universidade Federal da Paraíba

Centro de Informática

Programa de Pós-Graduação em Informática

Uma nova abordagem para a identificação de ilhas genômicas em bactérias com base no método de agrupamento *mean shift*

Daniel Miranda de Brito

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Informática da Universidade Federal da Paraíba como parte dos requisitos necessários para obtenção do grau de Mestre em Informática.

Orientadora: Thaís Gaudencio do Rêgo

João Pessoa - PB

Fevereiro de 2017

B862n Brito, Daniel Miranda de.
Uma nova abordagem para a identificação de ilhas genômicas em bactérias com base no método de agrupamento *mean shift* / Daniel Miranda de Brito.- João Pessoa, 2017.
74 f. : il.-

Orientadora: Thaís Gaudencio do Rêgo.
Dissertação (Mestrado) – UFPB/CI

1. Ilhas Genômicas. 2. Predição. 3. Análise Genômica. 4. Agrupamento. 5. *Mean Shift*. I. Título

UFPB/BC CDU – 004(043)

UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Ata da Sessão Pública de Defesa de Dissertação de Mestrado de DANIEL MIRANDA DE BRITO, candidato ao título de Mestre em Informática na Área de Sistemas de Computação, realizada em 24 de fevereiro de 2017.

Aos vinte e quatro dias do mês de fevereiro, do ano de dois mil e dezessete, às oito horas, no Centro de Informática da Universidade Federal da Paraíba, em Mangabeira, reuniram-se os membros da Banca Examinadora constituída para julgar o Trabalho Final do Sr. Daniel Miranda De Brito, vinculado a esta Universidade sob a matrícula nº 2015103447, candidato ao grau de Mestre em Informática, na área de "*Sistemas de Computação*", na linha de pesquisa *"Sinais, sistemas digitais e gráficos"*, do Programa de Pós-Graduação em Informática, da Universidade Federal da Paraíba. A comissão examinadora foi composta pelos professores: Thais Gaudêncio Do Rego (PPGI-UFPB), Orientadora e Presidente da Banca, Leonardo Vidal Batista (PPGI-UFPB), Examinador Interno, Sávio Torres De Farias (UFPB), examinador externo ao programa, e João Paulo Matos Santos Lima (UFRN), Examinador externo à instituição. Dando início aos trabalhos, a Presidente da Banca cumprimentou os presentes, comunicou aos mesmos a finalidade da reunião e passou a palavra ao candidato para que o mesmo fizesse a exposição oral do trabalho de dissertação intitulado "Uma nova abordagem para a identificação de ilhas genômicas em bactérias com base no método de agrupamento mean shift". Concluída a exposição, o candidato foi arguido pela Banca Examinadora que emitiu o seguinte parecer: "***aprovado***". Do ocorrido, eu, Clauirton de Albuquerque Siebra, Coordenador do Programa de Pós-Graduação em Informática, lavrei a presente ata que vai assinada por mim e pelos membros da banca examinadora. João Pessoa, 24 de fevereiro de 2017.

Prof. Dr. Clauirton de Albuquerque Siebra

Clauirton de Albuquerque Siebra
Coordenador do Programa de
Pós-Graduação em Informática
SIAPE 1723491

Prof. Drª. Thais Gaudêncio Do Rego
Orientadora (PPGI-UFPB)

Prof. Dr. Leonardo Vidal Batista
Examinador interno (PPGI-UFPB)

Prof. Dr. Savio Torres De Farias
Examinador externo ao programa (UFPB)

Prof. Dr. João Paulo Matos Santos Lima
Examinador externo à instituição (UFRN)

Agradecimentos

A Deus, por iluminar os meus caminhos e me dar forças para enfrentar todos os reveses que surgiram nessa longa caminhada.

À minha mãe Suerde Brito, meu pai João Brito e meu irmão Filipe Brito, pelo amor e apoio incondicionais, que seguramente foram imprescindíveis para essa vitória pessoal e profissional.

À minha querida namorada Mariana Rosal, por me ajudar a reencontrar o caminho da felicidade.

Aos meus avós, tios, primos, amigos e professores que me incentivaram e apoiaram durante todo o mestrado.

Aos meus amigos Eduardo Queiroga, Iron Araújo, Marcílio Lemos, Matheus Cordeiro e Rubens Correia, e aos amigos Leandro Paiva e Túlio Pascoal, por todo a consideração e amizade verdadeira, desde o início da minha graduação.

À minha amiga Thaís Ratis, pela disposição em ajudar e sempre contribuir com o desenvolvimento da pesquisa.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes), pelo importante papel no fomento da pesquisa científica no país.

Ao professor Dr. Lucídio Cabral, por ter me dado importantes conselhos em momentos de dúvida durante toda a trajetória do mestrado.

Ao professor Dr. Vinícius Maracaja-Coutinho, pela significativa colaboração para o desenvolvimento deste trabalho.

À professora Dra. Thaís Gaudencio, pelo privilégio de tê-la como orientadora e amiga. Sou grato por suas orientações, que muito acrescentaram aos meus conhecimentos acadêmicos, e por sua paciência e dedicação, determinantes para eu chegar até aqui.

Aos membros da banca examinadora, professores Dr. Leonardo Vidal, Dr. Sávio Farias e Dr. João Paulo Matos, por terem aceitado o convite para participar da banca e pelas relevantes contribuições dadas ao trabalho.

"A tarefa não é tanto ver aquilo que ninguém viu, mas pensar o que ninguém ainda pensou sobre aquilo que todo mundo vê."

Arthur Schopenhauer

Resumo

Ilhas genômicas (IGs) são regiões do genoma de bactérias e arqueas adquiridas por meio do fenômeno da transferência horizontal. Frequentemente, essas regiões proporcionam importantes adaptações a esses organismos, como resistência a antibióticos e patogenicidade, cujos efeitos podem ser danosos a outras espécies. Por essa razão, diversas metodologias computacionais foram propostas para a sua predição, porém nenhuma capaz de identificar o repertório completo de ilhas presentes em uma determinada sequência genômica. Portanto, torna-se oportuno o desenvolvimento de novas abordagens que explorem diferentes aspectos dessas regiões, permitindo a identificação daquelas não conhecidas. Nesse trabalho, propõe-se um novo método para a identificação de IGs, construído com base no algoritmo de agrupamento *mean shift*, com o cálculo automático da largura de banda, indispensável para o seu funcionamento. Resultados dos testes com ilhas genômicas inseridas em genomas de bactérias mostram que o método é capaz de identificar essas regiões com taxas de acerto acima de 99%. Testes realizados com genomas de bactérias com IGs conhecidas revelaram o potencial do método para a sua identificação e para a descoberta de novas ilhas. O estudo detalhado do conteúdo das novas ilhas apontou a presença de elementos típicos de IGs, confirmando a eficácia do método na predição dessas regiões.

Palavras-chave: Ilhas genômicas, predição, análise genômica, agrupamento, *mean shift*

Abstract

Genomic islands (GIs) are regions of the bacterial and archaeal genomes that were acquired through the phenomenon of horizontal transfer. Usually, these regions provide important adaptations to these organisms, such as antibiotic resistance and pathogenicity, whose effects can be harmful to other species. For these reasons, many computational methodologies have been proposed for their prediction, however, none of them are capable to precisely identify the whole repertoire of islands present in a given genomic sequence. Therefore, the development of new approaches that explore different aspects of these regions is timely, allowing the identification of those not known. In this paper, it is proposed a novel method for the identification of GIs, built based on mean shift clustering algorithm, with the automatic bandwidth calculation, necessary to its operation. Test results with genomic island inserted in bacterial genomes show that the method is capable of identifying these regions, with sensitivity rates above 99%. Tests performed with bacterial genomes with known GIs revealed the potential of the method for their identification and for the discovery of new island. The detailed study of the new islands content pointed the presence of typical GIs elements, confirming its effectiveness in the prediction of these regions.

Keywords: Genomic Islands, prediction, genomic analysis, clustering, mean shift

Conteúdo

Lista de Símbolos

bp : *Base pair, em inglês; ou Pares de bases, em português*

DNA : *Deoxyribonucleic Acid, em Inglês; ou Ácido Desoxirribonucleico, em Português*

IDE : *Integrated Development Environment, em inglês; ou Ambiente de Desenvolvimento Integrado, em português*

IG : *Ilha Genômica*

IM : *Ilha de Metabolismo*

IP : *Ilha de Patogenicidade*

IR : *Ilha de Resistência*

IS : *Ilha de Simbiose*

NCBI : *Centro Nacional de Informações Biotecnológicas, em inglês; ou Centro Nacional de Informações Biotecnológicas, em português*

RNA : *Ribonucleic Acid, em inglês; ou Ácido Ribonucleico, em português*

THG : *Transferência Horizontal de Genes*

TLG : *Transferência Lateral de Genes*

Lista de Figuras

Lista de Tabelas

Lista de Equações

Capítulo 1

Introdução

As bactérias constituem o maior domínio de vida do planeta Terra e suas espécies atuais ocupam uma enorme variedade de nichos ecológicos, apresentando impressionantes diversidades molecular, metabólica e ecológica, o que demonstra o enorme potencial evolutivo desses organismos [Langille et al. 2008; Hentschel e Hacker 2001]. O primeiro genoma bacteriano sequenciado por completo foi o da bactéria *Haemophilus influenzae*, em meados da década de 1990 [Fleischmann et al. 1995]. Desde então, com o enorme avanço das capacidades computacionais e tecnologias de sequenciamento, diversos outros genomas de bactérias e arqueas foram sequenciados e disponibilizados em bases de dados públicas. O avanço frequente dessas tecnologias indica que a quantidade de genomas acessíveis para estudo continuará a crescer com muita rapidez [Madigan et al. 2015].

A rápida coleta de informações de sequências biológicas impulsionou o desenvolvimento de algoritmos e recursos computacionais para acessar e analisar as informações disponíveis [Attwood et al. 2011]. O estudo dessas sequências genômicas proporcionou o entendimento de conceitos nas mais diversas áreas, tais como medicina e microbiologia [Madigan et al. 2015]. No campo da microbiologia, uma das importantes descobertas foi o fato das bactérias possuírem, além dos genes responsáveis por codificar suas funções metabólicas essenciais, uma quantidade variável de genes relacionados a funções acessórias, adquiridos de outros organismos, por intermédio do fenômeno da Transferência Horizontal de Genes (THG). Essas porções de DNA adquiridas formam blocos sintéticos conhecidos como Ilhas Genômicas (IGs) e destacam-se pelo seu papel crítico, sendo frequentemente responsáveis pela capacidade de adaptação à mudança de condições em ambientes clínicos, industriais e naturais

[Juhas et al. 2009]. Ademais, a THG desempenha função fundamental na evolução dos procariotos desde a sua origem [Doolittle 1999]. A importância do fenômeno no processo evolutivo das bactérias é verificada por meio da análise do genoma da bactéria *Escherichia coli*, que possui, segundo estimativas, cerca de 17% da sua composição genética proveniente de organismos de espécies diferentes, através da THG [Hacker e Kaper 2000], e da bactéria *Thermotoga maritima*, que presume-se ter mais de 24% dos seus genes provenientes de arqueas [Nelson et al. 1999].

As IGs são de grande interesse para a comunidade científica, visto que podem conter genes que codificam funções acessórias, como: resistência a antibióticos, patogenicidade, virulência e diferentes adaptações com efeitos adversos a outros organismos [Hacker e Kaper 2000]. Ainda no rol de adaptações conferidas às bactérias pelas ilhas genômicas estão a capacidade de degradação de compostos recalcitrantes e produção de metabólitos secundários [Che et al. 2014b]. Essas regiões, a depender do tempo em que ocorreu a transferência, são capazes de exibir diferenças significativas na organização da sua informação biológica, o que torna possível a sua identificação a partir da análise genômica [de Brito et al. 2016].

É fundamental prover meios de reconhecimento das IGs, porquanto a sua descoberta permite a investigação detalhada dos genes que as compõem, facilitando o entendimento das funções das bactérias e dos seus principais aspectos evolutivos [Che et al. 2014b]. A identificação das ilhas é uma importante etapa do desenvolvimento de novas vacinas e medicamentos para combater doenças. Tendo em vista a importância das regiões, diversos métodos foram propostos para a sua identificação, cada um considerando seus diferentes aspectos. A despeito da vasta disponibilidade de métodos, nenhum deles é capaz de identificar com precisão o repertório completo de IGs de uma determinada bactéria; o que se deve à grande variedade da distribuição de nucleotídeos nas diferentes espécies de bactérias. Torna-se, pois, oportuno o desenvolvimento de novos métodos, empregando diferentes abordagens, de forma a aumentar a eficiência preditiva na identificação das IGs, mediante a integração das múltiplas abordagens existentes [de Brito et al. 2016].

Do ponto de vista médico, é importante dispor de métodos robustos e confiáveis para identificar rapidamente ilhas genômicas em linhagens de bactérias virulentas recém sequenciadas, para as quais a necessidade de desenvolvimento de antibióticos pode ser urgente, conforme publicação da Organização Mundial de Saúde (OMS) [Organização Mundial de

Saúde 2017]. Conhecer as ilhas genômicas é o primeiro passo para a descoberta de alvos moleculares que podem ser utilizados no desenho racional de fármacos.

Neste trabalho, propõe-se uma nova abordagem para a identificação de IGs em bactérias, desenvolvida a partir do método de agrupamento *mean shift*. A sua eficiência é atestada com fundamento em dois experimentos distintos: o primeiro, considerando a inserção de IGs em um dado genoma bacteriano; e o segundo, avaliando o genoma de bactérias estudadas em outros trabalhos da área, em busca de ilhas genômicas novas e das já conhecidas. Busca-se, assim, que a abordagem proposta possa ser utilizada, individualmente ou em conjunto com os diversos métodos já existentes, para a identificação de IGs em bactérias, de forma que se possa usufruir do máximo benefício obtido com a identificação dessas regiões, como por exemplo a descoberta de vacinas e medicamentos candidatos.

1.1 Objetivos

1.1.1 Objetivo Geral

Desenvolver uma abordagem para a identificação de ilhas genômicas em organismos procariotos, fundamentada no método de agrupamento *mean shift*.

1.1.2 Objetivos Específicos

- Elaborar método para a predição de IGs em sequências de DNA de procariotos, com base em uma estratégia de agrupamento semissupervisionado com o *mean shift*;
- Estudar a viabilidade do método desenvolvido na identificação de IGs inseridas em genomas de bactérias;
- Verificar a capacidade do método em identificar novas regiões, ainda não descritas;
- Comparar a eficiência da abordagem com aquelas já existentes na literatura.

1.2 Publicações Relacionadas

[de Brito et al. 2015] de Brito, D. M., Ramos, T. A. R., Maracaja-Coutinho, V., de Farias,S. T., Batista, L. V., e do Rêgo, T. G. (2015). Using the mean shift clustering algorithm to predict genomic islands in bacteria. In *Proceedings X-Meeting 2015*.

[de Brito et al. 2016] de Brito, D. M., Maracaja-Coutinho, V., de Farias, S. T., Batista, L. V., e do Rêgo, T. G. (2016). A novel method to predict genomic islands based on mean shift clustering algorithm. *PloS one*, 11(1):e0146352.

1.3 Estrutura da Dissertação

O restante deste texto está organizado da seguinte forma:

Capítulo 2: Apresenta-se a fundamentação teórica necessária para o desenvolvimento do trabalho, conceitos de genoma, aprendizagem de máquina e agrupamento;

Capítulo 3: As principais classes de métodos de detecção de ilhas genômicas e os seus principais expoentes são apresentados;

Capítulo 4: Mostra-se a proposta da abordagem para a identificação de IGs e seus principais aspectos são discutidos;

Capítulo 5: A metodologia utilizada para a avaliação do modelo apresentado é exposta, em dois enfoques: bactérias com ilhas genômicas embutidas e bactérias com IGs conhecidas;

Capítulo 6: Os resultados obtidos na avaliação são apresentados e discutidos;

Capítulo 7: Exibem-se as considerações finais e os trabalhos futuros

Capítulo 2

Fundamentação Teórica

Este capítulo tem como objetivo apresentar uma revisão dos principais conteúdos necessários para o desenvolvimento desta pesquisa. Apresentam-se aspectos fundamentais da Biologia, desde a organização do genoma de um organismo, até a descrição do fenômeno da transferência horizontal de genes, responsável pela formação das ilhas genômicas. Também são apresentados conceitos da área de aprendizagem de máquina, com foco no aprendizado não-supervisionado e sua principal tarefa, o agrupamento, cujo detalhamento é dado ao seu representante *mean shift*, base da elaboração da proposta aqui apresentada.

2.1 Transferência Horizontal de Genes e Ilhas Genômicas

2.1.1 O genoma

O genoma é responsável por conter as informações hereditárias de cada organismo vivo, sendo mais de dez milhões de espécies diferentes existentes no planeta Terra [Alberts et al. 2015; Lewin 2004]. Ele carrega todas as informações genéticas do organismo e revela pistas do seu funcionamento e da sua história evolucionária [Madigan et al. 2015]. Pode-se pensar o genoma em duas perspectivas: a funcional e a física (estrutural). Na funcional, ele é composto por um conjunto de genes. Um gene representa uma unidade de informação genética, sendo cada um deles responsável por codificar uma proteína ou uma molécula de RNA [Lewin 2004].

Na perspectiva física, o genoma consiste de uma longa sequência de monômeros de ácido

nucléico, denominados nucleotídeos [Lewin 2004]. Um nucleotídeo, por sua vez, é composto de três componentes principais, quais sejam [Madigan et al. 2015]: pentose, bases nitrogenadas e grupo fosfato. A pentose é a desoxirribose para o ácido desoxirribonucleico (do inglês *deoxyribonucleic acid*, DNA) e a ribose para o ácido ribonucleico (do inglês *ribonucleic acid*, RNA). As bases nitrogenadas são: adenina (representada pela letra A), citosina (C) e guanina (G) para ambos DNA e RNA, timina (T) exclusivamente para o DNA, e uracila (U) para o RNA.

O genoma é composto pelo conjunto inteiro de cromossomos de um organismo, os quais se compõem de moléculas de DNA, que carregam as informações genéticas das células dos organismos. Os genes nele presentes servem de molde para a produção de uma outra sequência, o RNA. Esta é uma molécula intermediária cujas instruções são convertidas em diferentes proteínas. Além disso, é interessante citar que algumas moléculas de RNA não são traduzidas, mas possuem funções catalíticas, regulatórias ou estruturais [Madigan et al. 2015; Alberts et al. 2015]. O processo típico de conversão de uma sequência de DNA em uma proteína é observado na Figura 2.1.

A sequência de nucleotídeos de uma molécula de DNA ou RNA compõe a sua estrutura primária e constitui a informação genética de um organismo [Madigan et al. 2015]. A molécula de DNA possui normalmente duas fitas de ácido nucleico, unidas por ligações de hidrogênios formadas entre os nucleotídeos de cada uma delas [Madigan et al. 2015]. Para as ligações de hidrogênio, apenas pares de nucleotídeos A-T e C-G são formados, garantindo que as duas fitas de DNA sejam complementares [Madigan et al. 2015; Lewin 2004]. Dessa forma, é possível determinar qual o nucleotídeo presente na outra fita, conhecendo-se apenas aquele presente em uma das fitas.

Os genomas dos organismos apresentam uma grande variedade de tamanhos, podendo estender-se por centenas de milhares de nucleotídeos. Uma unidade típica para medição da extensão do genoma são os pares de base (do inglês *base pair*, bp), bem como os seus múltiplos: quilobase (kb), correspondendo a 1.000 bp; Megabase (Mb), representando 1.000.000 bp; e Gigabase (Gb), referindo-se a 1.000.000.000 bp. O menor genoma conhecido de um organismo de vida livre é o da bactéria *Mycoplasma genitalium*, que vive como parasita em mamíferos e possui apenas 530 genes, totalizando 580.470 bp (aproximadamente 580 kb) [Zhang e Zhang 2004b]. Porém existem relatos na literatura de bactérias simbiontes

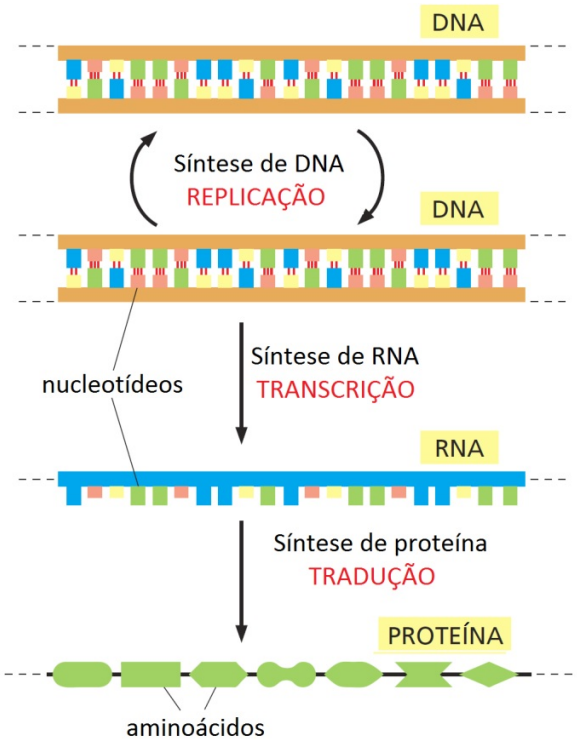

Figura 2.1: **Processo de transferência da informação genética, do DNA à proteína**

Na replicação, cada uma das fitas duplas de DNA são separadas e servem de molde para a síntese de novas fitas complementares, produzindo duas cópias idênticas do DNA com toda informação genética do organismo (replicação) [Alberts et al. 2015]. Em seguida, na transcrição, o DNA é utilizado como molde para a síntese do RNA, cujas instruções podem ser traduzidas em proteína (tradução), composta por uma sequência de aminoácidos.

Fonte: Traduzido de [Alberts et al. 2015].

com genomas menores, como é o caso da *Candidatus* nasuia deltocephalinicola, que possui um genoma de apenas 112 kb [Bennett et al. 2016]. Em um outro extremo, existe o genoma do ser humano com cerca de 30.000 genes, que por sua vez é composto por quase três bilhões de pares de nucleotídeos (aproximadamente 3,0 Gb) [Alberts et al. 2015; Venter et al. 2001].

O tamanho reduzido do genoma das bactérias facilitou a aplicação de técnicas de sequenciamento para determinar as suas sequências genômicas completas [Alberts et al. 2015]. No entanto, isso não é mais problema, tendo em vista que as técnicas mais atuais de sequenciamento permitem que sequências inteiras sejam determinadas em minutos ou horas, em vez de dias [van Dijk et al. 2014]. Sequenciamento refere-se ao processo bioquímico de deter-

minação da ordem precisa de nucleotídeos em uma sequência de DNA ou RNA [Madigan et al. 2015].

2.1.2 Transferência Horizontal de Genes

A transferência de informações genéticas entre células ocorre primariamente de modo vertical, isto é, o material genético é transferido da célula mãe para a célula filha durante o processo de divisão celular [Lawrence e Roth 1996]. No entanto, existe outra forma em que essa transferência pode ocorrer, qual seja, transferência horizontal de genes (THG), também conhecida como transferência lateral de genes (TLG). Na THG, uma porção de DNA pode ser transferida do genoma de um organismo para outro, independente de reprodução, mesmo que pertençam a espécies diferentes [Alberts et al. 2015].

A transferência horizontal entre organismos procariotos (bactérias e arqueas) é mediada por três processos (mecanismos) [Lawrence e Roth 1996; Alberts et al. 2015]: transformação, no qual um fragmento de DNA livre, liberado de um procarioto, é captado diretamente por um outro organismo naturalmente transformável (Figura 2.2 (a)); transdução, em que a transferência de DNA é feita por intermédio de um tipo de vírus bacteriano, denominado fago (Figura 2.2 (b)); e conjugação, quando a transferência de DNA requer interação célula-a-célula e um plasmídeo conjugativo na célula doadora (Figura 2.2 (c)).

O fenômeno da THG é muito comum em organismos procariotos, sendo de vital importância para a evolução dessas espécies [Alberts et al. 2015]. Em algumas bactérias, ele é responsável por grande parte das suas composições genéticas. Apesar de descrita com mais frequência nos organismos procariotos, a transferência horizontal também é percebida em organismos eucariotos superiores, com exemplos de transferências entre procariotos e eucariotos, e entre organismos eucariotos [Keeling e Palmer 2008; Andersson 2009; Marcet-Houben e Gabaldón 2010; Gladieux et al. 2014].

2.1.3 Ilhas Genômicas

Os fragmentos de DNA exógenos, transferidos horizontalmente de outros organismos, são comumente definidos como Ilhas Genômicas (IGs) e possuem tamanhos que variam tipicamente de 10 kb a 200 kb [Hacker e Kaper 2000]. Essas regiões tornaram-

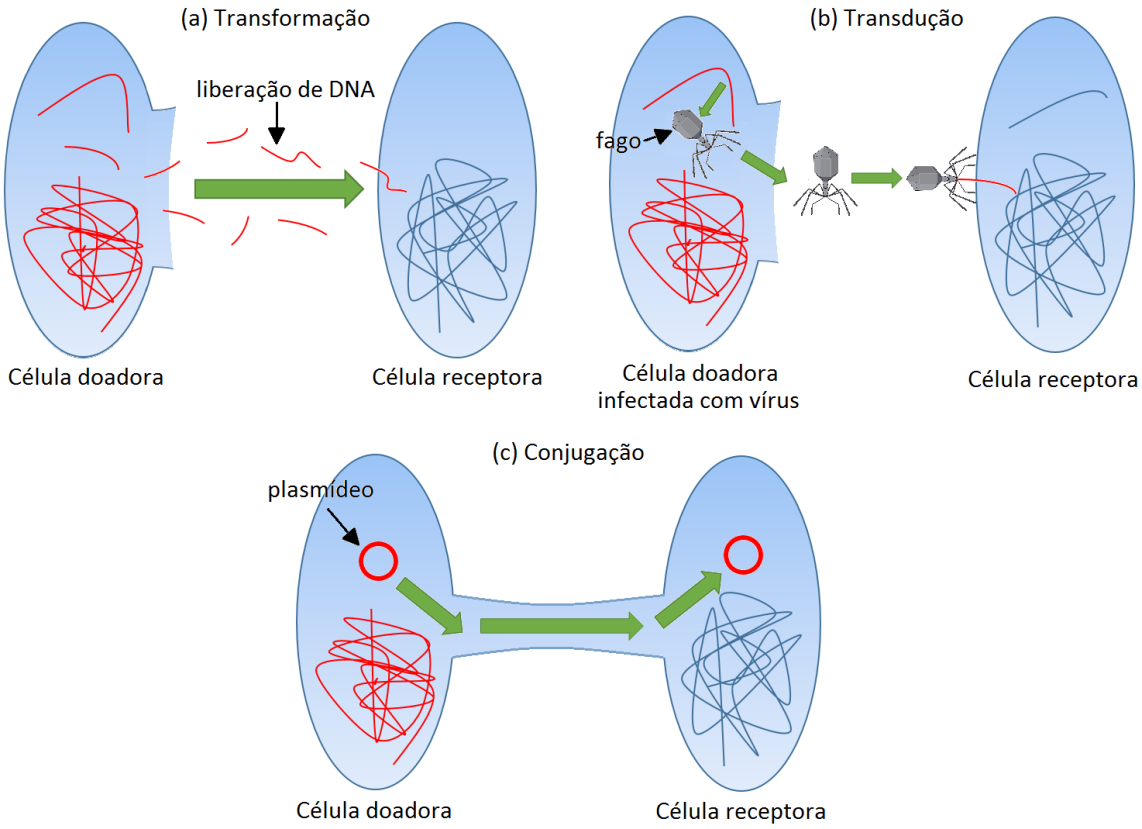

Figura 2.2: **Transferência horizontal de genes entre dois organismos**

(a) Na transformação, um fragmento de DNA liberado por um organismo é capturado por outro. **(b)** Na transdução, a transferência de DNA é mediada por um bacteriófago presente na célula doadora. **(c)** Na conjugação, o DNA é transferido por intermédio de um plasmídeo conjugativo presente na célula doadora e requer interação célula-a-célula.

Fonte: Elaborada pelo autor.

se de grande interesse científico, a julgar pela vasta quantidade de adaptações proporcionadas pelos genes que as compõem. É possível que elas estejam associadas à biossíntese de metabólitos secundários (Ilhas de metabolismo, IM), capacidade de simbiose com outros organismos (Ilhas de simbiose, IS), resistência a antibiótico (Ilhas de resistência, IR) e patogenicidade, em que fatores de virulência são carregados pelas ilhas, permitindo que os organismos causem doenças (Ilhas de patogenicidade, IP) [Mitić et al. 2008; Soares et al. 2015]. Tudo isso resulta em interesse médico, ambiental e industrial pelas IGs [Juhas et al. 2009].

Algumas características genômicas apresentadas pelas IGs permitem a sua identificação

em um determinado genoma receptor. Se comparadas com as partes restantes do genoma, as ilhas genômicas exibem assinaturas distintas, fato que é justificado pela origem em outros organismos [Che et al. 2014a]. A assinatura genômica de uma porção do genoma pode ser medida de diversas maneiras, sendo geralmente expressas pelo conteúdo G+C (%), frequência de oligonucleotídeos de tamanhos típicos de 2-9 nucleotídeos (conhecida como *k*-mers), sendo o uso de dinucleotídeos (2-mer) e códons (3-mer) os mais comuns [Langille et al. 2010; Che et al. 2014a].

Além da assinatura genômica, outros aspectos das IGs ajudam na sua descoberta, tal como a grande quantidade de genes codificadores de proteínas hipotéticas (que apresentam funções desconhecidas). Essa característica das IGs pode ser explicada pela observação de que boa parte dos doadores das ilhas ainda não foram cultivados, nem sequenciados, e as funções das regiões transferidas permanecem desconhecidas [Langille et al. 2010]. Outros elementos estão relacionados às IGs, sendo encontrados com frequências nas regiões, elencam-se: genes relacionados a fagos, transposases, integrases e genes associados à virulência [Che et al. 2014a].

O estudo das sequências genômicas é uma das áreas mais importantes da biologia computacional, no entanto, com o grande volume de dados acessíveis, encontra um importante desafio: transformá-los em conhecimento útil. As técnicas de aprendizagem de máquina são um aliado significativo no estudo e compreensão das sequências disponíveis [Larranaga et al. 2006].

2.2 Aprendizado de Máquina

Aprendizado (Aprendizagem) de Máquina é uma área da Ciência da Computação interessada no estudo de como fazer os computadores aprenderem a partir de dados [Han et al. 2011] e que busca construir programas inteligentes, capazes de serem aperfeiçoados automaticamente com base na experiência [Mitchell et al. 1997]. Fundada nessa experiência, elabora-se uma hipótese explicativa dos dados, permitindo que decisões futuras sejam tomadas segundo o conhecimento adquirido [Russell et al. 2011]. A área de aprendizado de máquina compreende, essencialmente, os problemas de aprendizado supervisionado, aprendizado não supervisionado e aprendizado semissupervisionado, os quais serão discorridos na

sequência.

2.2.1 Aprendizado supervisionado

Esta tarefa de aprendizagem de máquina busca encontrar, por meio da observação de um conjunto de pares entrada-saída, uma função que mapeie a entrada para a saída [Russell et al. 2011]. O aprendizado é dito supervisionado, pois um rótulo (saída) associado a cada tupla de treinamento é fornecido [Han et al. 2011]. Um exemplo de aprendizado supervisionado é a detecção de e-mails não solicitados (*spam*). Com base em um conjunto de documentos previamente especificados, um novo documento ainda não caracterizado é rotulado, a partir do seu conteúdo, como sendo legítimo ou *spam* [Guzella e Caminhas 2009].

Dado um conjunto de treinamento com *N* pares de exemplos entrada-saída: (x_1, y_1), (x_2, y_2), ..., (x_i, y_i), no qual cada y_i foi gerado por uma função desconhecida $y = f(x)$, objetiva-se encontrar uma função h que aproxime a função verdadeira [Russell et al. 2011]. Quanto ao valor da variável de saída, existem duas categorizações: diz-se haver um problema de **classificação**, quando a variável alvo pode assumir um conjunto finito de valores discretos (classes); caso a variável alvo seja contínua, trata-se de um problema de **regressão**. Na Figura 2.3 é possível ter a visão geral do processo de aprendizado supervisionado, em que um conjunto de treinamento é aplicado em um algoritmo de aprendizagem de máquina com intuito de se produzir um modelo descritivo dos dados (hipótese), que pode então ser utilizada para predizer a variável alvo de exemplos novos, dado o seu conjunto de atributos de entrada.

2.2.2 Aprendizado não-supervisionado

O aprendizado não-supervisionado visa, basicamente, encontrar padrões úteis nos dados. Ao contrário do aprendizado supervisionado, os rótulos associados a cada instância de treinamento não são conhecidos e o número de classes (grupos) presentes nos dados pode não ser determinado [Han et al. 2011]. O agrupamento é a sua principal tarefa. A descoberta de grupos de estudantes de acordo com as suas habilidades em uma determinada disciplina é um exemplo de aprendizado não-supervisionado [Merceron e Yacef 2005]. Os grupos são formados a partir das respostas providas pelos estudantes em exercícios e a sua identificação

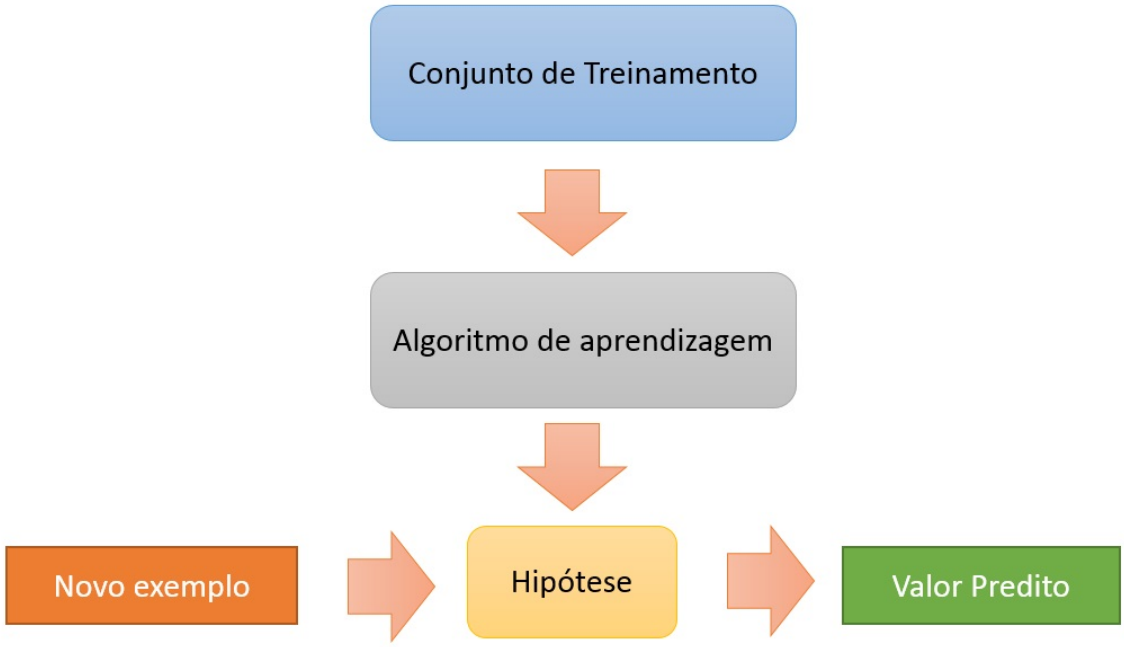

Figura 2.3: **Visão geral do aprendizado supervisionado**

Um algoritmo de aprendizagem de máquina é aplicado no conjunto de treinamento rotulado para a elaboração de uma hipótese explicativa dos dados, que permite que exemplos futuros sejam preditos a partir do modelo elaborado.

Fonte: Elaborada pelo autor.

permite que *feedbacks* sejam fornecidos aos professores, de modo a auxiliar a tomada de decisões pedagógicas.

Dada a importância da atividade de agrupamento para o desenvolvimento deste trabalho, discute-se sobre a mesma em detalhes em uma seção exclusiva.

2.2.3 Aprendizado semissupervisionado

É uma classe de técnicas de aprendizagem de máquina que utiliza exemplos rotulados e não rotulados para aprender um modelo [Han et al. 2011]. Para classificação, os exemplos não rotulados podem ser utilizados para refinar as fronteiras entre as classes. Inicialmente cria-se um modelo explicativo dos dados, utilizando os exemplos rotulados e, em seguida, o algoritmo tenta classificar as instâncias não rotuladas, que serão utilizadas posteriormente para a aprimorar o modelo criado [Han et al. 2011]. No agrupamento semissupervisionado,

utilizam-se rótulos fornecidos pelo usuário ou restrições para guiar o agrupamento na escolha de grupos mais apropriados [Basu et al. 2004].

2.3 Agrupamento

Agrupamento ou análise de *clusters* é o processo de particionamento de um conjunto de objetos (ou observações) em subconjuntos [Han et al. 2011]. Cada subconjunto é definido como um grupo ou *cluster*, no qual objetos pertencentes a um mesmo *cluster* são mais similares entre si do que aqueles pertencentes a *clusters* diferentes [Jain et al. 1999]. O seu propósito é encontrar informações ocultas presentes nos dados, ajudando na descoberta de grupos desconhecidos [Tsai et al. 2014]. Um grupo de objetos (*cluster*) pode ser tratado como uma classe implícita, por isso o processo de agrupamento é conhecido como **classificação automática**. A análise de *clusters* igualmente pode ser utilizada para a identificação de *outliers* (valores "isolados" de qualquer grupo). Nesses casos, a identificação dos objetos distantes pode ser mais importante do que os casos comuns [Han et al. 2011].

No agrupamento, o particionamento não é gerado por humanos, mas sim por um algoritmo de agrupamento. A aplicação de diferentes métodos de agrupamento pode produzir diferentes *clusters*, mesmo quando aplicados em um mesmo conjunto de dados. A análise de agrupamento possui aplicações em várias áreas, tais como: segmentação de imagens, visão computacional, reconhecimento de padrões e biologia computacional [Berkhin 2006a]. Uma atividade típica de agrupamento compreende os estágios apresentados na Figura 2.4 [Jain et al. 1999; Faceli et al. 2011]:

O primeiro estágio engloba as etapas de extração e seleção dos atributos dos objetos (padrões). Os atributos representam as suas caraterísticas e servem para distingui-los de outros objetos. Nesse estágio, pode ser necessário incluir uma etapa de preparação, que engloba normalizações, conversões de tipo e redução do número de atributos [Faceli et al. 2011]. A etapa seguinte abarca a definição da métrica (função) de distância entre quaisquer pares de objetos, necessária para medir a similaridade entre os mesmos. A definição da medida depende do domínio da aplicação e qual o tipo de informação que se deseja extrair [Faceli et al. 2011]. Uma das métricas mais utilizadas é a distância Euclidiana, definida na Equação 2.1

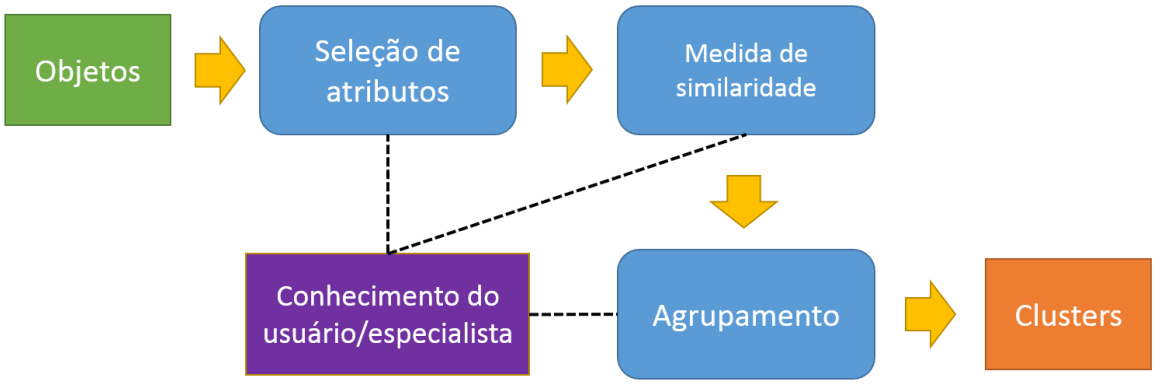

Figura 2.4: **Estágios do processo de agrupamento**

O processo de agrupamento tem o objetivo de dividir um conjunto de objetos não rotulados em grupos (*clusters*), de modo que padrões antes ocultos possam ser encontrados nos mesmos. As linhas tracejadas indicam que o conhecimento do especialista no domínio do problema estudado é requerido em todas as etapas do agrupamento.

Fonte: Elaborada pelo autor.

$$d(x_i, x_j) = \sqrt{\sum_{i=1}^{d} (x_i^l - x_j^l)^2} \tag{2.1}$$

O principal estágio do processo de agrupamento consiste em aplicar os algoritmos de agrupamento aos dados, para que sejam identificadas neles possíveis estruturas [Faceli et al. 2011]. Os grupos encontrados podem ser exclusivos, em que cada elemento pertence a exatamente um único *cluster* ou não-exclusivos, sendo cada elemento associado a um ou mais grupos [Jain e Dubes 1988]. No primeiro caso, a aplicação de um algoritmo exclusivo produz uma partição dos dados, isto é, um conjunto de *clusters* e seus respectivos membros, cuja definição formal é a seguinte: dado um conjunto de dados $P = \{p_1, p_2, ..., p_n\}$, no qual cada elemento p_i consiste de um vetor de d atributos, isto é, $p_i = (p_{i1}, p_{i2}, ..., p_{id})$, um algoritmo particional tenta encontrar uma partição $P = \{C_1, C_2, ..., C_k\}$, onde $k < n$ [Faceli et al. 2011; Han et al. 2011]. Uma partição pode ser definida como um conjunto de grupos e seus respectivos membros/elementos. Uma partição é resultado da aplicação de um algoritmo de agrupamento no conjunto de padrões. As partições devem manter as propriedades enumeradas a seguir:

1. Cada *cluster* deve conter pelo menos um elemento: $C_i \neq \emptyset$

2. Não deve haver elementos em comum entre dois ou mais *clusters*: $C_i \cap C_j = \emptyset$

3. Todo elemento deve estar associado a um *cluster*: $\bigcup_{i=1}^{k} C_i = P$

O conhecimento do usuário/especialista na aplicação pode ser utilizado em todas as etapas do processo de agrupamento, tendo papel importante para guiar a escolha das técnicas mais adequadas e na interpretação dos clusters obtidos. Ao término das etapas discutidas, pode-se fazer necessária a aplicação de etapas posteriores, a saber: validação e interpretação. A primeira é responsável por avaliar o resultado de um agrupamento em termos de um índice de validação, que pode ser externo, quando a estrutura produzida é comparada a uma estrutura previamente conhecida; interno, que tenta determinar se a estrutura encontrada é apropriada aos dados, considerando características intrínsecas dos dados; e relativo, que compara duas estruturas e mede o seu mérito relativo [Jain et al. 1999]. A última etapa refere-se ao processo de examinar cada grupo produzido em relação a seus objetos, de forma a descrever a sua natureza e atribuir significados e relações entre eles. Deve ser realizada por uma pessoa que detenha conhecimento no domínio [Faceli et al. 2011].

Duas são as principais categorizações de um algoritmo de agrupamento exclusivo, a hierárquica e a particional [Jain et al. 1999; Berkhin 2006a]. No agrupamento hierárquico, produz-se como saída uma sequência de partições aninhadas. Os métodos hierárquicos subdividem-se em aglomerativos (*bottom-up*) e divisíveis (*top-down*), de acordo com o modo em que a hierarquia é formada [Han et al. 2011]. Na primeira abordagem, cada elemento do conjunto inicia em seu próprio *cluster*, que é recursivamente combinado com os *clusters* mais apropriados, enquanto que na segunda abordagem, todos os elementos pertencem inicialmente ao mesmo *cluster*, e vão sendo recursivamente divididos em grupos cada vez menores. Algoritmos hierárquicos têm a vantagem de não requerer a definição prévia do número de *clusters* e o seu resultado é o mesmo independente da inicialização. O fato do mesmo ser estático, isto é, pontos atribuídos a um mesmo *cluster* não podem ser movidos para outros *clusters*, é uma desvantagem, assim como a dificuldade de separação de *clusters* sobrepostos, devido a falta de informação sobre a forma global ou tamanho dos *clusters* [Frigui et al. 1999; Das et al. 2008]. Embora não requeira a especificação do número de *clusters*, é necessário

definir um "ponto de corte" na hierarquia de partições produzida pelo algoritmo, o que, muitas vezes, é algo difícil de se fazer [Jung et al. 2003].

Por outro lado, o agrupamento particional tenta decompor o conjunto de dados diretamente em um conjunto de *clusters* disjuntos. Funções objetivo são usadas como critério de otimização durante o particionamento dos dados [Jain et al. 1999]. A desvantagem dessa abordagem é a necessidade de especificar o número de grupos *a priori* na maior parte dos algoritmos, o que nem sempre é possível, principalmente quando se lida com conjuntos de dados desconhecidos [Tsai et al. 2013], e a sensibilidade a ruído e *outliers* [Frigui et al. 1999]. Métodos que usam essa abordagem são relativamente escaláveis e simples, além de serem adequados para a descoberta de *clusters* compactos e bem separados [Sisodia et al. 2012].

Duas outras categorizações dos métodos de agrupamento são utilizadas na prática: os métodos baseados em densidade e os métodos baseados em *grid*. Os primeiros baseiam-se na ideia geral de que *clusters* são regiões densas do espaço de objetos, separadas por regiões esparsas, que crescem em qualquer direção [Berkhin 2006b]. São adequados para a descoberta de grupos de formas arbitrárias, além de permitirem que os *outliers* sejam filtrados [Han et al. 2011]. Os últimos quantizam o espaço de objetos em um conjunto finito de células que formam uma estrutura de *grid*. Apresentam como vantagem um rápido tempo de processamento [Han et al. 2011]. Um determinado algoritmo de agrupamento pode integrar as várias ideias apresentadas, o que impossibilita, em grande parte das vezes, classificá-lo unicamente em uma das categorias expostas [Han et al. 2011].

2.3.1 *Mean Shift*

O *mean shift* é um algoritmo de agrupamento particional, baseado em densidade, que não exige que o número de grupos presentes nos dados seja especificado *a priori*. O método visa encontrar as modas presentes nos dados através de uma rotina de convergência [Cheng 1995; Georgescu et al. 2003]. Aspectos relevantes, como garantia de convergência, robustez e facilidade de implementação tornaram-no popular em vários campos de aplicação [Sasaki et al. 2014].

Matematicamente, o *mean shift* é definido como segue: Dado um conjunto de *n* pontos $\mathbf{x}_i$ no espaço d-dimensional, onde $i = 1, ..., n$, a estimação de densidade multivariada com

kernel $K(\mathbf{x})$, calculada no ponto $\mathbf{x}$ pode ser definida da seguinte maneira:

$$\hat{f}(\mathbf{x}) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right) \tag{2.2}$$

onde *h* denomina-se *bandwidth* (em português, largura de banda) e define o tamanho do *kernel*

A função *kernel* é definida como $K(\mathbf{x}) = c_{k,d}k(\|\mathbf{x}\|^2)$, onde $c_{k,d}$ é a constante de normalização que integra $K(\mathbf{x})$ a 1, e $k(x)$ é denominado perfil do *kernel*. Calculando o gradiente do estimador de densidade definido em Eq. (2.2), temos a equação abaixo, depois de alguma manipulação algébrica, assumindo a função, $g(x) = -k'(x)$:

$$\hat{f}(\mathbf{x}) = \underbrace{\frac{2c_{k,d}}{nh^{d+2}} \left[\sum_{i=1}^{n} g\left(\left\|\frac{\mathbf{x} - \mathbf{x}_i}{h}\right\|^2\right)\right]}_{1^{\text{o}} \text{ termo}} \underbrace{\left[\frac{\sum_{i=1}^{n} \mathbf{x}_i g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)} - \mathbf{x}\right]}_{2^{\text{o}} \text{ termo}} \tag{2.3}$$

O primeiro termo da Eq. (2.3) é proporcional a Eq. (2.2), e o segundo termo é o vetor *mean shift* $\mathbf{m}_h(\mathbf{x})$, descrito abaixo:

$$\mathbf{m}_h(\mathbf{x}) = \frac{\sum_{i=1}^{n} \mathbf{x}_i g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)} - \mathbf{x} \tag{2.4}$$

O vetor *mean shfit* sempre aponta na direção de aumento máximo da densidade. O procedimento realizado para um dado ponto $\mathbf{x}_i$ é descrito da seguinte forma:

1. Calcula-se o vetor *mean shift* $\mathbf{m}_h(\mathbf{x}_i^t)$;
2. Move-se a janela de estimação de densidade $\mathbf{x}_i^{t+i} = \mathbf{x}_i^t + \mathbf{m}_h(\mathbf{x}_i^t)$;
3. Repetem-se os passos acima até a convergência, isto é, quando, $\mathbf{x}_i^{t+1} - \mathbf{x}_i^t < \epsilon$.

onde o sobrescrito *t* é a iteração atual do procedimento, e ϵ, é o limiar. A aplicação do procedimento para um dado ponto $\mathbf{x}_i$ é ilustrada na Figura 2.5, onde os quadrados coloridos definem os sucessivos centros da janela de estimação, as setas vermelhas definem o vetor de deslocamento *mean shift*, os pontos vermelhos representam os dados de entrada para o

cálculo dos centroides e os círculos pontilhados denotam as janelas de estimação utilizadas até a convergência ser alcançada na *n*-ésima iteração.

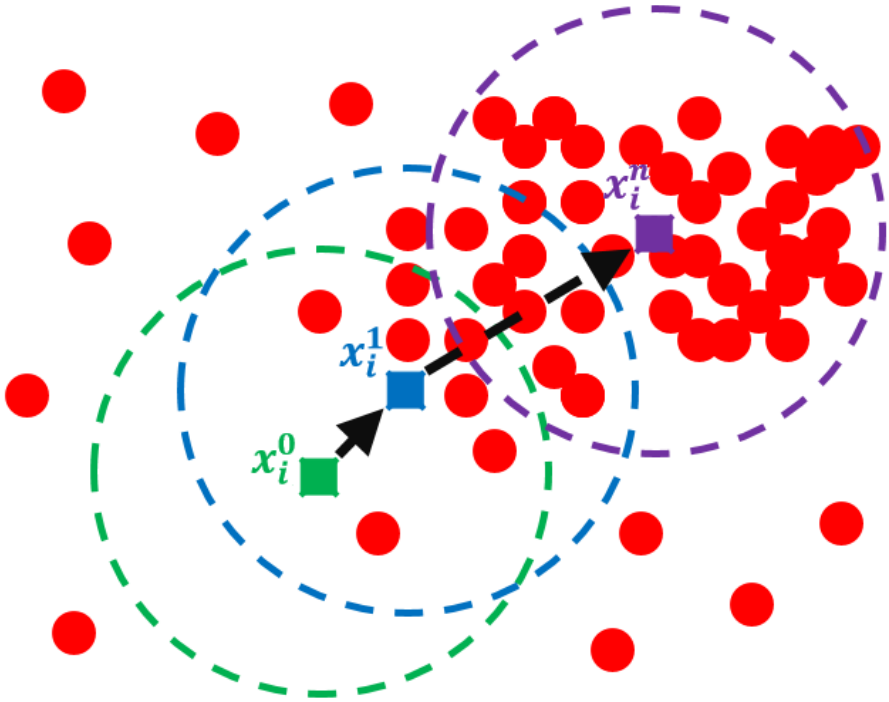

Figura 2.5: **Aplicação do *Mean shift* em um dado ponto** $\mathbf{x}_i$

Os pontos vermelhos representam os dados de entrada para a cálculo do ponto estacionário, os quadrados coloridos definem os sucessivos centros das janelas de estimação, as setas vermelhas associadas definem o vetor *mean shift*, os círculos coloridos pontilhados representam a janela utilizada na estimação da densidade até a convergência ser alcançada na *n*-ésima iteração.

Fonte: Elaborada pelo autor.

Para o agrupamento, o procedimento descrito deve ser aplicado em cada ponto do conjunto de dados. Pontos que convergirem para a mesma moda determinam um único *cluster*. A moda associada a cada ponto é obtida através do deslocamento sucessivo do conjunto de dados para a média dos pontos da sua vizinhança (centroide), até a convergência. O número de *clusters* é definido automaticamente, através das modas descobertas [Georgescu et al. 2003].

Capítulo 3

Trabalhos Relacionados

Neste capítulo, são apresentados alguns dos principais trabalhos desenvolvidos para a identificação de ilhas genômicas em organismos procariotos. Os trabalhos dividem-se, sobretudo, em duas abordagens: os que se baseiam em composição de sequência (homologia independente), isto é, utilizam apenas as características inerentes à sequência para a predição das regiões divergentes; e os métodos baseados em comparação genômica (homologia dependente), em que uma ou mais sequências são utilizadas como base para a predição de ilhas em um dado genoma investigado. Também são discutidas as ferramentas que integram diversos métodos de predição.

3.1 Métodos baseados em composição de sequência

A observação de que os genes de uma espécie particular de bactéria usualmente compartilham uma mesma assinatura genômica é a base de construção de grande parte dos métodos de predição de IGs. Nessa abordagem, as regiões do genoma que exibem assinaturas diferentes são consideradas IGs. Os critérios utilizados para capturar a assinatura genômica incluem conteúdo G+C, frequência de dinucleotídeos e *codon usage* [Che et al. 2014a]. Para que os métodos possam ser aplicados, é comum dividir o genoma em fragmentos e compará-los com os valores esperados para uma dada medida. Esses métodos são ditos “baseados em composição de sequência". Dentre os métodos que utilizam essa abordagem destacam-se: SIGI-HMM [Waack et al. 2006], que se fundamenta na análise de *codon usage* de cada gene do genoma estudado para a identificação de IGs, bem como a da sua provável origem.

Compara-se o *codon usage* de cada gene analisado com uma tabela de *codon usage* representando bactérias doadoras ou genes diferencialmente expressos. O método realiza múltiplos testes para identificar os potenciais genes exógenos, incorporando o Modelo Oculto de Markov (do inglês *Hidden Markov Model*, HMM) para otimizar a predição das regiões. PAI-IDA [Tu e Ding 2003] utiliza análise discriminante iterativa para definir regiões incomuns no genoma de uma dada bactéria, de acordo com três critérios de composição: conteúdo G+C, frequência de dinucleotídeos e *codon usage*. Essas regiões são consideradas IGs. O Island-Path [Hsiao et al. 2003] utiliza múltiplos sinais de DNA e anotações com características do genoma para predição e visualização gráfica de IGs identificadas. As unidades básicas de identificação das ilhas são conteúdo G+C das janelas preditas (do inglês *open reading frame*, ORF), frequência de dinucleotídeos para *clusters* de genes, localização dos prováveis elementos genéticos móveis e localização dos tRNA.

O Centroid [Rajan et al. 2007] é outro método baseado em composição de sequência. Primeiramente, ele particiona o genoma em regiões de tamanhos iguais e representa a frequência de distribuição de bases de cada região em pontos no espaço *n*-dimensional, cuja dimensão é determinada pela fórmula $n = 4^m$, a qual representa o número de palavras distintas de tamanho *m* formadas pelos símbolos {A,T,C,G}. O centroide de todos os pontos é calculado e utilizado como critério para determinar os *outliers* entre eles, que são considerados IGs.

A ideia do método INDeGenIUS [Shrivastava et al. 2010] é similar à do Centroid. Ele utiliza o princípio do agrupamento hierárquico. Inicialmente, divide o genoma em fragmentos de tamanhos iguais e para um dado oligonucleotídeo de tamanho *k*, o vetor de frequência de cada 4^k palavras é computado. Em uma etapa seguinte, *n clusters* são criados e a matriz de distância entre os *clusters* é calculada. Com base nessa matriz, o par com menor distância é agrupado em um único *cluster*. Calcula-se a porcentagem de elementos de cada *cluster* e repete-se o processo até que o número de elementos do *cluster* majoritário (com maior número de elementos) seja menor do que um dado limiar. Terminado o processo de formação dos grupos, calcula-se o centroide do *cluster* majoritário, utilizado para encontrar as IGs.

Igualmente com base em composição de sequência, destaca-se o software PIPS [Soares et al. 2012], que se concentra na identificação de ilhas de patogenicidade. Ele é capaz de utilizar múltiplas características das IPs de uma maneira integrada para a sua identificação.

Posteriormente, os autores do PIPS constataram a ausência de métodos para a identificação de outros tipos de IGs (IM, IR e IS) e estenderam o seu primeiro método para a predição dessas classes de ilhas, em um novo software denominado GIPSy [Soares et al. 2015].

3.2 Métodos baseados em comparação genômica

Outro grupo de métodos para identificação de IGs baseia-se em comparação genômica. Parte-se do pressuposto de que genomas de espécies estreitamente relacionadas devem compartilhar preferências genômicas e assinaturas similares. Portanto, se uma espécie genômica contém assinaturas não presentes nas outras espécies, é altamente aconselhável assumir que estas sequências possuam origem em outros organismos [Che et al. 2014a]. Os métodos baseados nesta abordagem selecionam genomas de espécies filogeneticamente relacionadas a um dado genoma investigado, alinhando-os ao genoma estudado. Fragmentos que não compartilham as assinaturas genômicas presentes nas outras espécies são considerados IGs. Na sequência são descritos dois dos principais métodos desta categoria.

IslandPick [Langille et al. 2008] é um método baseado em comparação genômica que seleciona automaticamente genomas relacionados para comparação, utilizando uma matriz de distância pré-computada para medir o grau de semelhança com o genoma de referência e, em seguida, faz o alinhamento múltiplo do genoma consultado com aqueles selecionados na etapa anterior. A partir do alinhamento, extraem-se as regiões únicas do genoma investigado, que são consideradas IGs. MobilomeFinder [Ou et al. 2007] é outro método que compara sequências próximas para identificar IGs em um dado genoma. A sua ideia é semelhante a empregada pelo IslandPick, porém o processo de seleção de sequências relacionadas é feito manualmente e é limitado a identificação de IGs iniciadas por tRNA (*transfer* RNA), classe de moléculas de RNA na qual muitas IGs são inseridas [Alberts et al. 2015; Che et al. 2014a].

3.3 Ferramentas e recursos para predição de IGs

Algumas ferramentas disponibilizam a opção de utilizar múltiplos métodos para predição de IGs, permitindo a visualização e comparação dos resultados obtidos. EGID [Che et al.

2011] é um *ensemble* de cinco métodos de detecção de IGs: Alien_Hunter, SIGI-HMM, INDeGenIUS, IslandPath e PAI-IDA. A ferramenta recebe como entrada as sequências genômicas e anotações de genes, e utiliza um esquema de votação para definir as potenciais IGs, a partir dos resultados produzidos pelos métodos implementados. GIST [Hasan et al. 2012] é outra ferramenta de *ensemble*, que implementa os mesmos métodos do EGID, além de prover uma versão otimizada do mesmo. Os autores propuseram uma ferramenta com interface amigável e facilidade de fazer *download* dos genomas diretamente da base do NCBI. O IslandViewer [Dhillon et al. 2015] é uma aplicação Web com interface amigável para obter IGs pré-computadas ou realizar novas predições a partir dos métodos IslandPick, IslandPath-DIMOB e SIGI-HMM. Ainda permite visualização fácil e download em vários formatos das IGs identificadas.

3.4 Discussão

Tanto a abordagem baseada em composição de sequência, quanto a de comparação genômica apresentam vantagens e desvantagens. As abordagens baseadas em composição de sequência têm como vantagem utilizar apenas o genoma investigado, tornando possível a predição de IGs em todas sequências genômicas [Langille et al. 2008]. Contudo, métodos baseados nessa abordagem podem levar a predições de falsas IGs, devido a presença de sequências anormais no genoma, como genes altamente expressos [Langille et al. 2008], e deixar de identificar IGs que foram adquiridas de genomas com composição de sequência similar ou IGs incorporadas ao genoma que sofreram o efeito de *amelioration*, no qual a assinatura genômica do genoma doador se ajusta a do genoma residente [Langille et al. 2008; Lawrence e Ochman 1998].

A vantagem da abordagem baseada em comparação genômica é a facilidade de identificar diferenças entre sequências genômicas diretamente relacionadas. A desvantagem da abordagem baseada em comparação genômica é a necessidade de uso de genomas de espécies filogeneticamente próximas para comparação, o que nem sempre é possível, principalmente quando genomas de bactérias próximas ainda são desconhecidos. Outra desvantagem é que a maior parte dos métodos requer ajuste manual de parâmetros, que é difícil de ser realizado e pode levar a inconsistências no resultado das predições [Che et al. 2014a].

A Tabela 3.1 apresenta um resumo esquemático dos principais trabalhos relacionados, apontando a abordagem utilizada em cada método e as suas principais características.

Tabela 3.1: Resumo esquemático dos principais trabalhos relacionados

Trabalho	**Abordagem utilizada**	**Descrição do método**
SIGI-HMM [Waack et al. 2006]	Composição de sequência	Prevê IGs e o provável doador de cada gene individual, a partir da análise de *codon usage* de cada gene do genoma analisado.
PAI-IDA [Tu e Ding 2003]	Composição de sequência	Utiliza a análise discriminante para definir regiões que desviam do resto do genoma, em três critérios: conteúdo G+C, frequência de dinucleotídeos e *codon usage*.
IslandPath [Hsiao et al. 2003]	Composição de sequência	Incorpora múltiplos sinais de DNA e anotações com características dos genomas para predizer IGs. As características incluem conteúdo G+C, dinculeotídeos, localização dos elementos genéticos móveis e dos tRNAs.
Centroid [Rajan et al. 2007]	Composição de sequência	Particiona o genoma em regiões de tamanhos iguais e calcula a composição de sequência de cada região. O centroide de todas as regiões é calculado e usado como critério para identificação das IGs.
INDeGenIUS [Shrivastava et al. 2010]	Composição de sequência	Divide o genoma em grupos de tamanhos iguais e inicia o processo de agrupamento hierárquico, que prossegue até que o *cluster* majoritário incorpore um determinado número de elementos. Terminada a etapa, o centroide do grupo maior é calculado e usado como base para a identificação das IGs.
PIPS [Soares et al. 2012]	Composição de sequência	Desenvolvido para a identificação de ilhas de patogenicidade, utiliza diversas características das IPs de maneira integrada.

GIPSy [Soares et al. 2015]	Composição de sequência	Construído a partir do PIPS, foi extendido para identificar outros tipos de ilhas genômicas: IM, IR e IS.
IslandPick [Langille et al. 2008]	Comparação genômica	Seleciona automaticamente os genomas utilizados para comparação, em seguida faz o alinhamento múltiplo deles com o genoma estudado. A partir do alinhamento, extraem-se as regiões únicas do genoma investigado, que são consideradas IGs.
Mobilome-Finder [Ou et al. 2007]	Comparação genômica	Ideia similar à do IslandPick, diferenciando-se porque faz a seleção dos genomas de forma manual e limita-se a identificar IGs iniciadas por tRNA.
EGID [Che et al. 2011]	*Ensemble* de ferramentas	Implementa cinco métodos distintos e utiliza um esquema de votação entre eles para identificar as ilhas genômicas.
GIST [Hasan et al. 2012]	*Ensemble* de ferramentas	Apresenta os mesmos métodos do EGID, porém implementados de forma otimizada, com facilidades extras, como download de genomas direto da base de dados do NCBI.
IslandViewer [Dhillon et al. 2015]	*Ensemble* de ferramentas	Interface web amigável que contém um conjunto de IGs pré-computadas para consulta e ainda permite realizar novas predições a partir de vários métodos.

A abordagem proposta neste trabalho integra a categoria dos métodos baseados em composição de sequência, com algumas particularidades no seu funcionamento. Além do uso do genoma a ser analisado, empregam-se sequências artificiais na etapa de ajuste de parâmetros do método, construídas por meio de um paradigma de modelagem de DNA. O processo de seleção dessas sequências difere do empregado nos métodos de comparação genômica, uma vez que não exige que as sequências sejam relacionadas ao genoma investigado. Todas as características do método são discutidas em detalhes no Capítulo 4.

Capítulo 4

Predição de Ilhas Genômicas com o *Mean Shift*

Neste capítulo, apresenta-se a proposta de uma nova abordagem para a identificação de ilhas genômicas em bactérias: o *Mean Shift Genomic Island Predictor* (MSGIP), fundamentado no algoritmo de agrupamento *mean shift*. Os principais aspectos teóricos do método são apontados, bem como discussão acerca da sua capacidade de identificação das IGs.

4.1 Abordagem proposta para a identificação das ilhas genômicas

O MSGIP apoia-se no fato de que IGs usualmente exibem características distintas do restante dos genes do DNA de uma bactéria, sendo a aplicação de um algoritmo de agrupamento capaz de identificar os *clusters* que desviam do restante do genoma (*outliers*), que são considerados IGs. O método é dito baseado em composição de sequência, haja vista utilizar-se apenas de informações da própria sequência investigada, não realizando comparações com sequências relacionadas.

Emprega-se no método a combinação de classificadores independentes (conhecida como *ensemble* ou comitê) para compor a predição final das IGs. Pretende-se, assim, reduzir os possíveis erros de classificações individuais, devido a aleatoriedade existente no processo de seleção dos fragmentos artificiais, e torná-lo mais determinístico, isto é, que nas suas várias

execuções, os resultados apresentados sejam mais parecidos entre si. Na abordagem proposta, são realizadas múltiplas classificações e as decisões individuais de cada classificador são combinadas em um único resultado, através de um esquema de votação.

A estratégia de combinação utilizada é a de votação não ponderada, ou seja, cada classificação tem o mesmo peso na definição do resultado final, que é decidido com base na coleção de todos os votos. A classificação de um fragmento como potencial ilha genômica ou região nativa do genoma é definida de acordo com a moda dos votos, isto é, a classe mais votada. Utilizam-se números ímpares como tamanhos de *ensembles* para evitar a ocorrência de empate e necessidade de um critério adicional para realizar o desempate. A Figura 4.1 ilustra o processo de combinação dos classificadores para produção do resultado final.

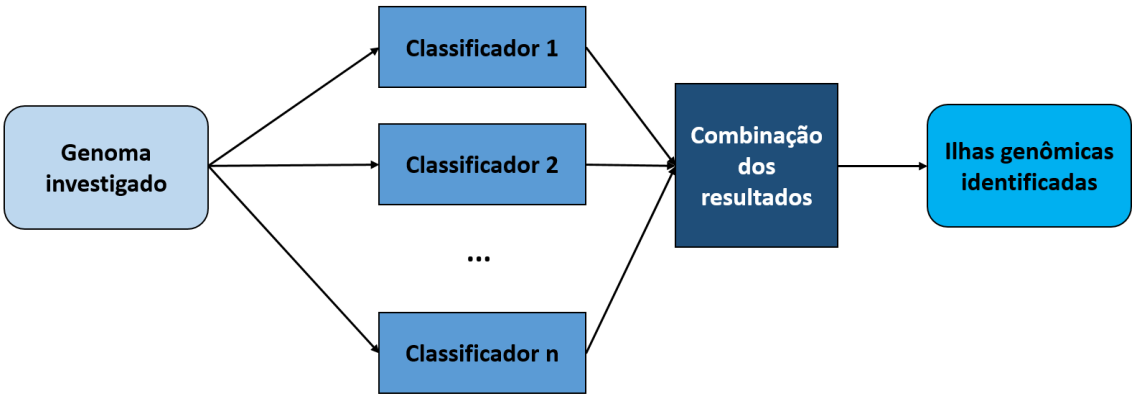

Figura 4.1: **Execução múltipla do MSGIP**

O genoma investigado é analisado de modo isolado pelos *n* classificadores distintos, os quais têm os resultados combinados mediante um esquema de votação não ponderada, que origina o resultado final a ser exibido ao usuário.

Fonte: Elaborada pelo autor.

Para realizar a predição, o método recebe como entrada um arquivo no formato FASTA *.fna*, que representa o genoma completo de uma determinada bactéria; o tamanho das unidades básicas de predição, aqui chamadas de fragmentos; tamanho *k* das palavras utilizadas para capturar a composição da sequência; e o tamanho do comitê de classificações a serem combinadas no resultado final. Produz como saída o conjunto de fragmentos que potencialmente hospedam ilhas genômicas. O fluxograma da abordagem proposta é mostrado na Figura 4.2.

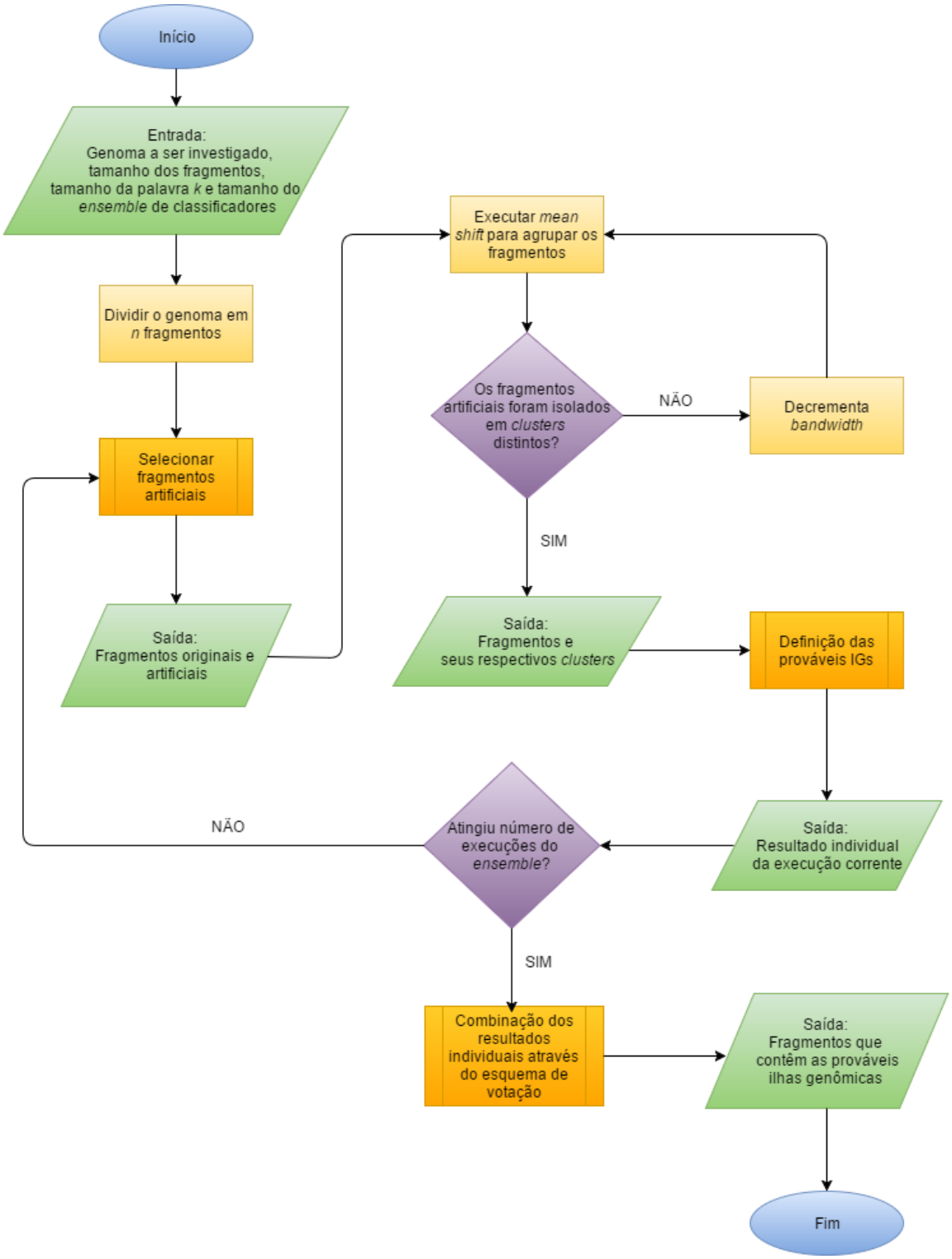

Figura 4.2: **Fluxograma da abordagem proposta**

O fluxograma inicia com a entrada do genoma investigado e dos parâmetros requeridos pelo MSGIP. Etapas intermediárias são necessárias para as execuções independentes de cada classificador e combinação dos resultados, que origina como saída os fragmentos que contêm as potenciais IGs.

Fonte: Elaborada pelo autor.

4.2 Classificações independentes das ilhas genômicas

O método proposto emprega uma coleção de classificadores para a composição do resultado final a ser exibido ao usuário. Inicialmente, o MSGIP particiona o genoma em *n* fragmentos não sobrepostos de tamanhos iguais. Em seguida, para cada classificador do *ensemble*, são selecionados fragmentos artificiais distintos com finalidade de determinar a *bandwidth* ótima (o critério de seleção utilizado é descrito com mais detalhes nas seções seguintes). A diferença entre os classificadores está nas sequências artificiais selecionadas e consequentemente no valor resultante da *bandwidth*. Com os fragmentos artificiais selecionados, o *mean shift* é aplicado sucessivas vezes para agrupar o conjunto formado pelos fragmentos originais e artificiais, até que o parâmetro *bandwidth* ótimo seja selecionado. Utiliza-se aqui o conceito de agrupamento semissupervisionado, haja vista que os fragmentos artificiais servem de "guia" para definição da largura de banda ótima que produza o agrupamento que melhor se ajuste aos dados.

Durante o processo de agrupamento, verifica-se a possibilidade de combinar os grupos produzidos pelo algoritmo, caso a distância (métrica empregada é a Euclidiana) entre eles seja menor do que o valor de $bandwidth/4$. Assim, espera-se evitar que um número muito elevado de *clusters* seja produzido, dificultando a identificação das IGs. O procedimento de convergência do *mean shift* se encerra quando o deslocamento da sua janela de convergência for menor que o limiar 10^{-3}. Ao término do processo de agrupamento, obtém-se uma lista de *clusters* e seus respectivos membros (fragmentos), cuja análise define as prováveis IGs.

Na análise dos agrupamentos, verifica-se se os fragmentos inseridos artificialmente estão isolados no seu próprio *cluster*. Caso estejam, encerra-se a execução do *mean shift*, caso contrário, repete-se o processo a partir da etapa anterior, executando o *mean shift* com *bandwidth* menor. Os *clusters* resultantes são investigados com o objetivo de identificar aqueles isolados dos fragmentos nativos do genoma (*outliers*), que são classificados como potenciais IGs, caso estejam presentes em agrupamentos de até 200 kb, valor definido com base no tamanho máximo típico encontrado em IGs, conforme discutido por [Hacker e Kaper 2000]. Todo o processo descrito é repetido o número de vezes indicado pelo tamanho do *ensemble*. O Algoritmo 1 mostra o pseudocódigo das classificações realizadas por cada membro do comitê.

Algoritmo 1: *Classificação realizada por cada classificador do ensemble*

Data: Lista dos fragmentos originais *lo* do genoma, tamanho de palavra *k*

Result: Possíveis IGs e suas respectivas posições no genoma

Lista IGs $\leftarrow \emptyset$;

Seleciona *m* fragmentos artificiais a partir do **Algoritmo 2**;

`/* Quanto maior for o valor de MAX_BANDWIDTH, maior será o número de iterações do algoritmo */`

bandwidth $\leftarrow$ *MAX_BANDWIDTH*

while ***true*** **do**

  agrupa fragmentos originais e artificiais com o *mean shift*;

  if *fragmentos artificiais foram separados* **then**

    Sair **while**;

  else

    `/* O valor de DEC_BANDWIDTH indica o quanto será decrescido da bandwidth para a iteração seguinte */`

    bandwidth $\leftarrow$ *bandwidth* $-$ *DEC_BANDWIDTH*

  end

end

for *cada fragmento original* **do**

  if *fragmento está em cluster com no máximo* $200kb$ **then**

    adiciona fragmento em *Lista IGs*

  end

end

return *Lista IGs*

4.2.1 Definição da *bandwidth* ótima

Apesar do *mean shift* ser não paramétrico, ele tem como principal limitação requerer que o parâmetro *bandwidth* seja especificado. Esse parâmetro determina o número de pontos da vizinhança no processo iterativo de deslocamento da janela de estimação e sua definição é crítica, pois afeta fortemente a quantidade de grupos formados e a qualidade do agrupamento [Wu e Yang 2007]. Por essa razão, define-se aqui a largura de banda global ótima automaticamente por meio de uma estratégia de agrupamento semissupervisionado, construída a partir da inserção de fragmentos artificiais no conjunto de fragmentos originais da bactéria e execução múltipla do *mean shift*. A restrição do agrupamento é que os fragmentos inseridos artificialmente sejam separados em *clusters* isolados. O processo de agrupamento é repetido com valores de *bandwidth* decrescentes, até que a restrição de separação seja satisfeita. Os fragmentos artificiais são produzidos aleatoriamente, de acordo com a estratégia definida nas Seção 4.2.2.

O conteúdo G+C é utilizado para calcular a dissimilaridade entre os fragmentos originais do genoma e os fragmentos candidatos (aqui denominada dissimilaridade interfragmentos) e entre os fragmentos artificiais já selecionados e os fragmentos candidatos (dissimilaridade intrafragmentos), conforme o Algoritmo 2. Há precaução em relação à dissimilaridade de conteúdo G+C entre os fragmentos, pois, caso ela seja muito baixa (indicando alta similaridade entre os fragmentos), é provável que o valor resultante de *bandwidth* seja muito pequeno, ocasionando a geração de um número muito grande de *clusters* pelo *mean shift*, dificultando a investigação dos agrupamentos formados na busca de IGs. Após experimentos com variações de parâmetros, definiram-se os valores mínimos de dissimilaridade para a seleção das sequências artificiais mais adequadas. Para fragmentos de 25 kb e palavras de tamanho $k = 4$, configuração utilizada em todos os experimentos descritos no Capítulo 5, os valores definidos para dissimilaridade interfragmentos e intrafragmentos foram de 2% e 3%, respectivamente.

A quantidade máxima de fragmentos *m* a serem selecionados pelo Algoritmo 2 depende do tamanho do genoma investigado. Considera-se o valor máximo de três fragmentos artificiais para genomas de até 0,65 Mb, cinco fragmentos para genomas de até 1,0 Mb, dez fragmentos para genomas de até 1,5 Mb e quinze fragmentos para genomas com tamanhos a partir de 1,5 Mb. O número de tentativas para a geração das sequências (*MAX_ITERAÇÕES*)

Algoritmo 2: Seleção dos fragmentos artificiais

Data: Lista *lo* de fragmentos originais, *min* dissimilaridade interfragmentos, *min* dissimilaridade intrafragmentos, número *m* de fragmentos artificiais

Result: Lista de fragmentos selecionados

```
1  listaFragmentosArtificiais ← ∅; cFragmentosSelecionados ← 0 ; numIterações ← 0
2  flagSimilaridade ← False
3  while cFragmentosSelecionados < m ∧ numIterações < MAX_ITERAÇÕES do
4      listaProbabilidades ← selecionaProbabilidades()
5      fragmentoCandidato ← constróiFragmento(listaProbabilidades)
6      gcCandidato ← calculaG+C(fragmentoCandidato)
7      for cada fragmento original o em lo do
8          gcOriginal ← calculaG+C(fragmento original o)
9          diferençaGC ← dissimilaridade(gcOriginal, gcCandidato)
10         if diferençaGC < mínima dissimilaridade interfragmentos then
11             flagSimilaridade ← True
12         end
13     end
14     if listaFragmentosArtificiais ≠ ∅;
15      then
16         for cada fragmento artificial a em listaFragmentosArtificiais do
17             gcArtificial ← calculaG+C(a)
18             diferençaGC ← dissimilaridade(gcArtificial, gcCandidato)
19             if diferençaGC <mínima distância intrafragmentos  then
20                 flagSimilaridade ← True
21             end
22         end
23     end
24     if flagSimilaridade = false then
25         adiciona fragmentoCandidato a listaFragmentosArtificiais;
26         cFragmentosSelecionados ← cFragmentosSelecionados + 1
27     end
28     numIterações ← numIterações + 1
29 end
30 return listaFragmentosArtificiais
```

é limitado em 1000, caso contrário, haveria possibilidade de o método executar indefinidamente por não conseguir selecionar todos os fragmentos de acordo com os critérios estabelecidos.

4.2.2 Geração dos fragmentos artificiais de DNA

Para evitar a necessidade do método incluir bases de dados com arquivos de genomas para a seleção dos fragmentos artificiais e consequente aumento do espaço gasto para mantê-lo em disco, utiliza-se aqui uma estratégia para produção desses fragmentos, a qual se baseia no modelo multinomial de sequências de DNA. O modelo pressupõe que as sequências são geradas por um processo estocástico que escreve aleatoriamente um dos quatro símbolos do alfabeto genético (A, T, C, G) em cada posição *i* do arquivo da sequência. Ele é especificado através da distribuição de probabilidade do alfabeto [Cristianini e Hahn 2006].

Para produzir fragmentos sintéticos com diversidades significativas de conteúdo G+C, a distribuição da probabilidade de frequência dos símbolos dos nucleotídeos é variável, de acordo com as observadas em diversas sequências reais (obtidas do RefSeq do NCBI). As probabilidades são guardadas em um arquivo de texto e a diversidade é alcançada por intermédio da seleção aleatória de uma das probabilidades salvas no arquivo para a construção das sequências artificiais.

4.2.3 Preenchimento automático dos fragmentos de fronteira

Devido ao fato do método proposto utilizar o critério de particionamento de fragmentos sem sobreposição, é possível que informações importantes de um determinado genoma sejam perdidas, caso se encontrem na fronteira final da sequência. É adequado, portanto, disponibilizar uma forma de manter os fragmentos localizados nas fronteiras. Isso é alcançado através do preenchimento automático do fragmento incompleto com a quantidade de bases remanescentes para completar o fragmento. Utilizam-se, para tal, as bases localizadas no começo do genoma. Isso é possível pelo fato da maior parte dos DNAs de organismos procariotos serem circulares [Alberts et al. 2015].

Capítulo 5

Avaliação Experimental

Neste capítulo, expõe-se a abordagem metodológica utilizada para avaliar o método desenvolvido. O MSGIP foi testado em diversas situações distintas, de modo que se pudesse atestar a qualidade das suas predições nos vários cenários práticos possíveis. Discutem-se as tecnologias utilizadas para a implementação e o plano de cada um dos experimentos, além da configuração do método utilizada em cada um deles. O resultado dos experimentos é exibido no Capítulo 6.

5.1 Ferramentas e Tecnologia

A abordagem proposta no trabalho foi desenvolvida na linguagem de programação Java, versão 1.8, com o apoio da IDE Eclipse, na sua versão Mars[1]. O método recebe como entrada um genoma no formato FASTA[2], na sua variante *.fna*, que contém o genoma completo de uma dada bactéria. Os arquivos dos genomas utilizados em todos os testes foram obtidos a partir da base de dados *Reference Sequence* (RefSeq), do NCBI (ftp://ftp.ncbi.nih.gov/genomes/archive/old_refseq/Bacteria/). Um script em Java foi desenvolvido para fazer o download das sequências, que são manipuladas de maneiras distintas, de acordo com a particularidade de cada experimento.

[1] https://eclipse.org/mars/

[2] http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml

5.2 Plano do Experimento

Para examinar a capacidade do método em encontrar as regiões exógenas presentes em uma sequência bacteriana, são propostos experimentos distintos: o primeiro elaborado para avaliar a sua sensibilidade na predição de ilhas genômicas embutidas em bactérias e o segundo planejado para investigar a sua coerência na análise de um conjunto de bactérias com IGs previamente descritas na literatura. Para incluir diversas circunstâncias práticas ao primeiro experimento, foi feita a sua subdivisão em experimentos menores, que tratam de pontos de discussão específicos, conforme detalhamento apresentado em seguida. Em todos os casos, uma classificação é considerada correta se o método for capaz de identificar o fragmento que contém a ilha genômica embutida. Todos os experimentos foram executados com tamanhos de fragmentos de 25 kb, palavras de $k = 4$ e tamanho de *ensemble* fixado em 5 membros. O conjunto de parâmetros foi definido com base em experimentos preliminares, comparando-se diferentes valores de tamanho de fragmentos (10 kb, 25 kb e 50 kb), de *k* (entre 1 e 5) e de número de membros do comitê (3 e 5). Um resumo de todos os experimentos e seus objetivos é exibido na sequência.

- **Detecção das ilhas genômicas embutidas nos genomas**: O MSGIP é avaliado a partir da investigação de sequências quiméricas com IGs embutidas. Tem como objetivo verificar a sensibilidade do método na identificação das regiões exógenas presentes no DNA.

- **Efeito da diferença do conteúdo G+C entre doador e receptor**: Com base na análise de sequências quiméricas com IGs de diferentes dissimilaridades de G+C entre a ilha e a sequência, visa encontrar a dissimilaridade mínima requerida pelo MSGIP que permita a identificação das ilhas genômicas.

- **Preenchimento dos fragmentos incompletos das fronteiras**: A estratégia proposta de manutenção dos fragmentos de fronteira é analisada. Tem o objetivo de confirmar a eficiência do método na predição de IGs em fragmentos incompletos.

- **Relação do tamanho dos fragmentos e das ilhas genômicas**: O experimento em questão visa investigar qual o impacto da relação entre o tamanho da IG buscada e

aquele definido para o fragmento. Utiliza-se como dados de entrada, os resultados obtidos no experimento da detecção de IGs embutidas nos genomas.

- **Bactérias com IGs conhecidas**: Nesta análise, são estudadas cinco bactérias em um maior nível de profundidade. Espera-se que o método encontre as IGs descritas em outros trabalhos e também novas IGs.

5.3 Bactérias com IGs embutidas

5.3.1 Detecção das ilhas genômicas embutidas nos genomas

Neste experimento, o método MSGIP é avaliado quanto à sua habilidade de encontrar fragmentos de DNA exógenos inseridos em uma dada sequência genômica. Os resultados obtidos por ele são comparados com obtidos pelos métodos tradicionais de predição: G+C e Assinatura Genômica [Karlin 2001], aplicados em um conjunto fixo de fragmentos, sem sobreposição. A taxa de acerto na identificação das IGs embutidas (sensibilidade ou taxa de verdadeiros positivos) é utilizada para avaliar o desempenho dos métodos. Devido a impossibilidade de considerar a especificidade (taxa de verdadeiros negativos) na análise dos resultados, por não ser viável identificar para cada sequência se uma região é uma provável ilha ou fragmento nativo, utiliza-se o percentual médio de ilhas genômicas encontradas nas sequências genômicas como possível indicativo do equilíbrio entre a identificação das IGs verdadeiras e não identificação dos fragmentos originais do genoma.

O conjunto de testes foi construído a partir das sequências obtidas do NCBI e é constituído apenas por bactérias com conteúdo G+C entre 25% e 75% (intervalo inclusivo), valores estes que compreendem a grande maioria das bactérias sequenciadas [Lightfield et al. 2011]. Replicou-se o conjunto para a construção de duas bases iguais, utilizadas para a seleção dos pares receptores/doadores e construção das sequências quiméricas empregadas no experimento. O processo inicia-se com a seleção do genoma receptor, utilizado como modelo para recepção da ilha genômica, que é feita por meio de amostragem aleatória simples sem reposição. O genoma doador é selecionado a partir de amostragem aleatória simples com reposição (é feita uma verificação para evitar que sejam selecionados receptores e doadores idênticos). A construção é feita através da substituição do conteúdo de uma região da bactéria receptora pelo conteúdo extraído da bactéria doadora. Todo o processo descrito é resumido na Figura 5.1.

Construíram-se sequências quiméricas com inserções coincidentes de ilhas genômicas de tamanhos 10 kb e 25. A inserção é dita coincidente quando o conteúdo da IG é posicionado entre o início e o final de um único fragmento. O conteúdo das sequências receptoras foi substituído pela região de posição correspondente das sequências doadoras (de tamanho igual a IG inserida) iniciada na posição 100 kb e terminada na posição (100 + tamanho da ilha) kb. Também foram definidas sequências quiméricas com IGs não coincidentes, isto é, cujo conteúdo inserido se divide em mais de um fragmento. Para esse tipo de inserção, utilizaram-se IGs de tamanho 25 kb e 50 kb extraídas e inseridas na região de 112,5 - 137,5 kb e 100 - 150 kb. Ao todo foram construídas 160 sequências para cada um dos tamanhos de IG em ambos os tipos de inserção.

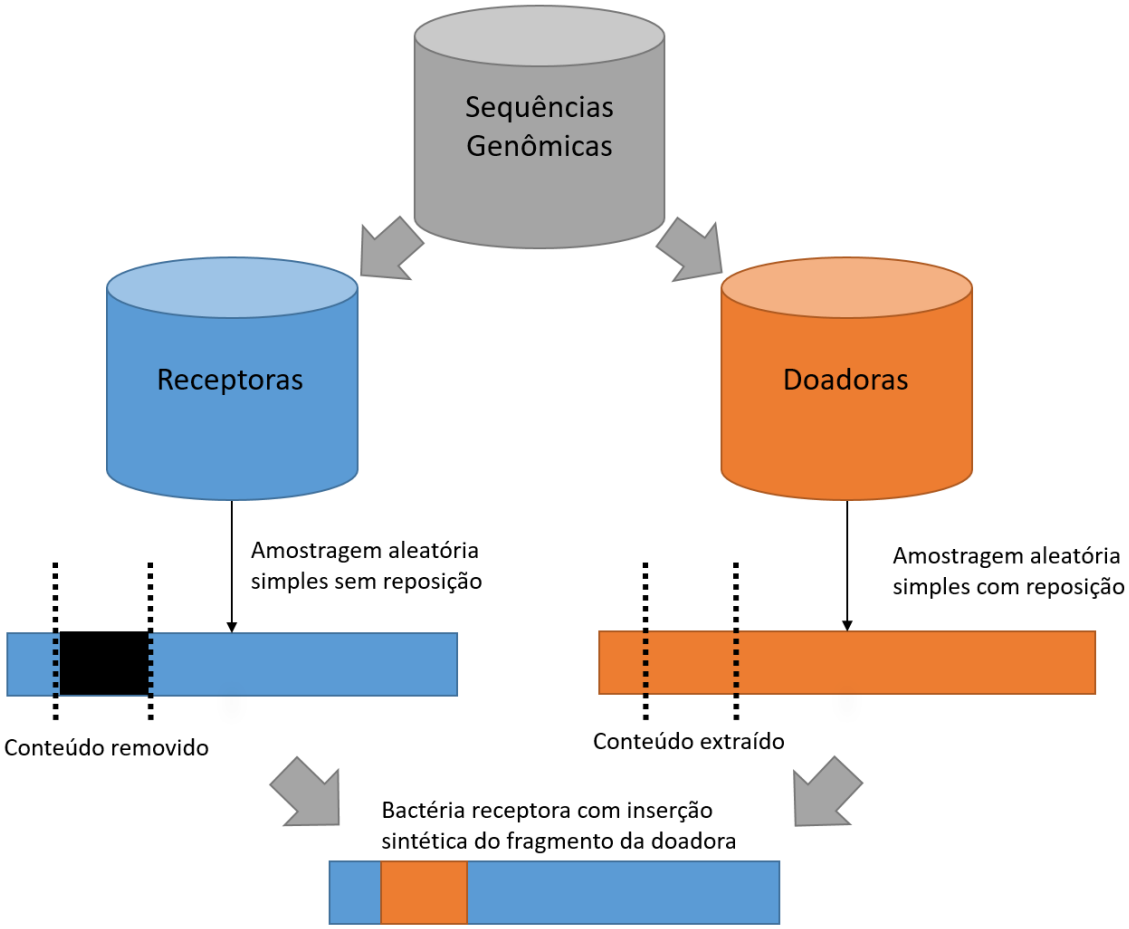

Figura 5.1: **Processo de confecção das sequências com IGs embutidas**

As sequências quiméricas foram construídas com base em duas sequências genômicas (receptora e doadora), selecionadas por amostragem aleatória simples. Para a sua confecção, o conteúdo da sequência receptora em uma posição fixa foi substituído pelo conteúdo extraído da sequência doadora na posição correspondente.

Fonte: Elaborada pelo autor.

5.3.2 Efeito da diferença do conteúdo G+C entre doador e receptor

Este experimento visa estudar a influência da similaridade entre o conteúdo G+C das IGs e dos genomas receptores na eficiência da predição das ilhas. Analisa-se a dissimilaridade mínima requerida pelo método para que a identificação das regiões divergentes seja possível. Do mesmo modo do experimento anterior, utiliza-se nesta análise sequências com conteúdo G+C entre 25% e 75%. O conjunto resultante foi subdividido em 10 subconjuntos menores, de acordo com o conteúdo G+C de cada sequência, tendo cada um deles uma variação de 5% em relação aos subconjuntos anterior e próximo. O processo de divisão descrito é ilustrado na Figura 5.2.

Para a aplicação desta análise, foram construídas sequências quiméricas com diferentes variações de conteúdo G+C entre receptor e doador. O processo de construção de cada sequência foi o seguinte: seleciona-se uma sequência aleatória do subconjunto "A", que é testada em relação às sequências presentes nos demais subconjuntos (selecionadas aleatoriamente), intencionando-se avaliar o impacto da diferença, que começa em 5% (diferença entre "A"e "B") e vai até 50% (diferença entre "A"e "J"). O processo de substituição das regiões é similar ao do experimento anterior, em que o conteúdo do genoma receptor é substituído pelo conteúdo do doador selecionado em uma posição fixa. Para cada valor de dissimilaridade, construíram-se 10 sequências com pares doadores/receptores distintos e ilhas genômicas de tamanho 25 kb.

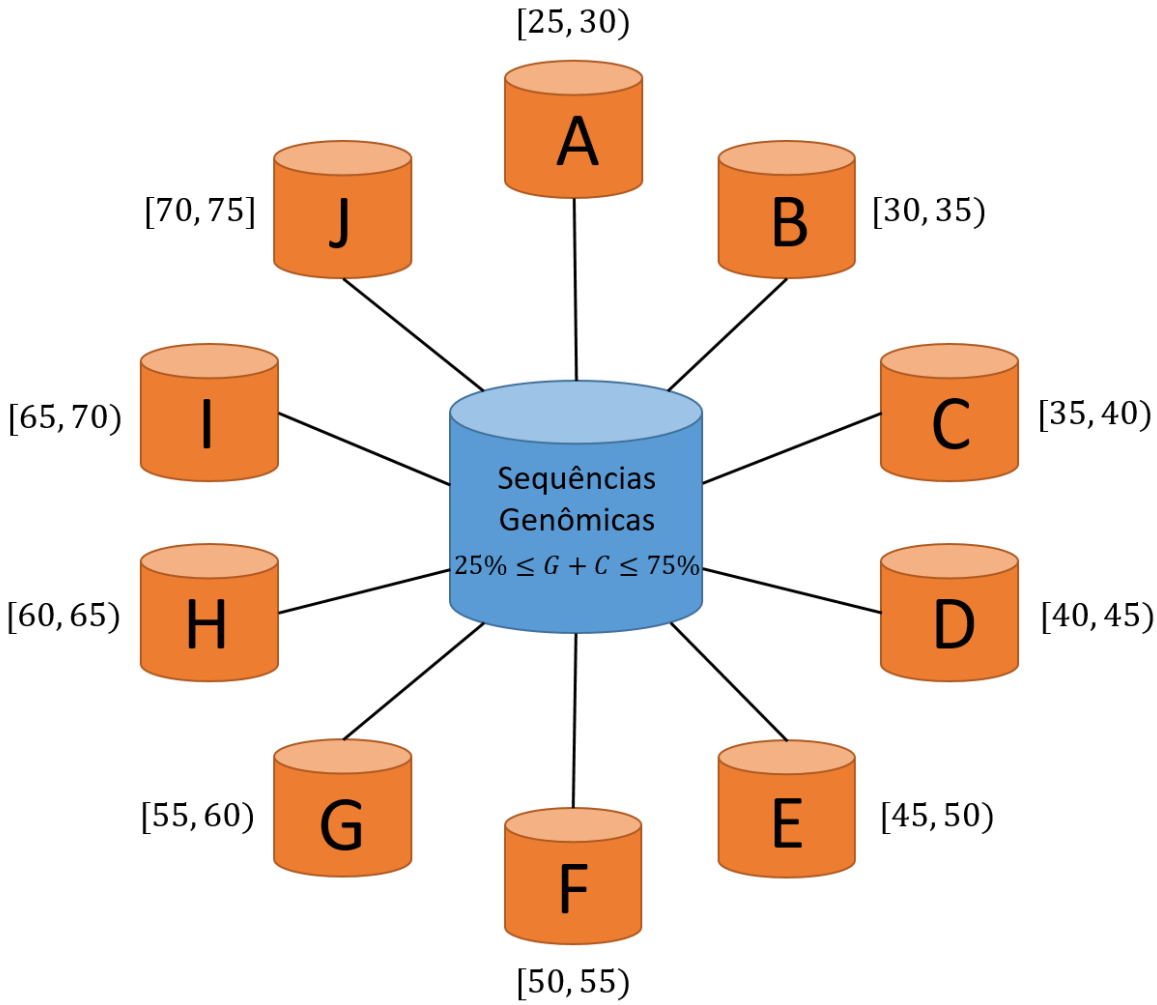

Figura 5.2: **Processo de divisão em subconjuntos de acordo com o conteúdo G+C**

Para analisar a dissimilaridade mínima de G+C (entre o genoma e a IG) requerida pelo MSGIP para que a identificação das ilhas seja possível, divide-se o conjunto de sequências genômicas com conteúdo G+C entre 25% e 75% em subconjuntos com variações de 5% entre os subconjuntos anterior e o próximo.

Fonte: Elaborada pelo autor.

5.3.3 Preenchimento dos fragmentos incompletos das fronteiras

O preenchimento automático dos fragmentos localizados na fronteira final das sequências genômicas é importante para evitar que informações relevantes neles contidas sejam perdidas, no caso de serem menores do que o valor definido para divisão do genoma. O experimento para analisar a eficácia desse recurso é proposto aqui.

Para tal, são construídas sequências quiméricas com ilhas genômicas incompletas embutidas na fronteira final do genoma. O que é feito através da substituição do fragmento original incompleto por um fragmento exógeno de tamanho pré-definido. Os receptores são selecionados por amostragem aleatória sem reposição; os doadores, por amostragem com reposição. São construídas ao todo 50 sequências, com inserções de 20%, 50% e 80% do tamanho do fragmento, para simular os vários casos possíveis. Utiliza-se tamanho de fragmentos de 25 kb, portanto os tamanhos das inserções são de 5 kb, 12,5 kb e 20 kb.

5.3.4 Relação do tamanho dos fragmentos e das ilhas genômicas

A presente análise visa investigar o impacto do tamanho dos fragmentos na sensibilidade do método na identificação de ilhas genômicas de diferentes tamanhos. Utiliza-se como base para este estudo, os resultados obtidos nos testes do experimento da Subseção 5.3.1. Utiliza-se um único tamanho de fragmento e três tamanhos diferentes de ilhas genômicas na investigação da proporção do tamanho da ilha genômica que se busca encontrar e do valor definido para o tamanho dos fragmentos.

5.4 Bactérias com IGs conhecidas

Genomas de bactérias com IGs previamente descritas na literatura são utilizadas neste experimento para a validação do método apresentado no trabalho. Uma vez que se busca estudar a capacidade do método na predição de novas IGs ainda não encontradas por outros métodos, as regiões encontradas pelo MSGIP que não foram descritas previamente na literatura são investigadas na busca de elementos típicos de IGs, de modo a confirmar a efetividade das predições. Os resultados obtidos com a aplicação do método em cada bactéria investigada, e suas discussões, são detalhados no capítulo seguinte. As bactérias utilizadas

neste trabalho foram selecionadas por amostragem não probabilística por julgamento e foram escolhidas por terem sido discutidas em artigos disponíveis na *web*. Elas são exibidas na Tabela 5.1.

Tabela 5.1: Bactérias analisadas pelo MSGIP

Bactéria	**Tamanho**	**Código**
Vibrio cholerae chromosome II	1,07 Mb	NC_002505.1
Corynebacterium glutamicum ATCC 13032	3,30 Mb	NC_003450.3
Streptococcus mutans UA159	2,03 Mb	NC_004350.2
Streptococcus pneumoniae G54	2,08 Mb	NC_011072.1
Rhodopseudomonas palustris CGA009	5,45 Mb	NC_005296.1

Capítulo 6

Resultados e Discussão

Neste capítulo, exibem-se os resultados dos experimentos descritos no Capítulo 5, realizados com sequências quiméricas incorporando IGs e com bactérias que contêm IGs conhecidas, os quais são analisados e discutidos na sequência.

6.1 Bactérias com IGs embutidas

6.1.1 Detecção das ilhas genômicas embutidas nos genomas

Os resultados obtidos no experimento pelo MSGIP e métodos relacionados são apresentados na Tabela 6.1. Por meio de sua análise, constata-se que IGs maiores têm melhores chances de identificação, sendo a de 25 kb, com a inserção coincidente, a que possui maior probabilidade de ser identificada. Verifica-se a superioridade do MSGIP na identificação das IGs de tamanhos 25 kb (coincidente), com sensibilidade de 99,3% e empate com o método Assinatura Genômica na identificação das ilhas de 50 kb, com 97,5% de acerto em ambos. Para a identificação de IGs de 10 kb, o método %G+C produziu resultados mais satisfatórios, com taxa de acerto de 78,6%, embora o MSGIP tenha gerado resultados igualmente satisfatórios (71,3% de acerto). Para a sequência quimérica com IG de 25 kb inserida de modo não coincidente, o método %G+C exibe os melhores resultados. No entanto, os métodos MSGIP e Assinatura Genômica apresentam resultados também satisfatórios.

Observa-se que os métodos tradicionais apresentam um elevado percentual médio de IGs preditas por genoma, o que indica que, apesar da boa sensibilidade por eles proporcionada,

existe a possibilidade dos mesmos identificarem um grande número de falsos positivos, o que compromete a eficiência dos métodos em questão, tornando o MSGIP presumivelmente a opção com maior equilíbrio entre as métricas sensibilidade e especificidade.

Tabela 6.1: Resultados dos testes com as sequências com IGs embutidas

Tamanho da IG	**Método**	**Sensibilidade**	**Percentual médio de IGs**
10 kb	MSGIP ($k = 4$)	71,3%	6,6%
	%G+C	78,6%	22,4%
	Assinatura Genômica	70,0%	14,1%
25 kb (coincidente)	MSGIP ($k = 4$)	99,3%	7,9%
	%G+C	91,3%	17,4%
	Assinatura Genômica	98,8%	13,0%
25 kb (não coincidente)	MSGIP ($k = 4$)	78,9%	9,7%
	%G+C	83,8%	21,5%
	Assinatura Genômica	78,1%	13,1%
50 kb	MSGIP ($k = 4$)	97,5%	10,7%
	%G+C	85,6%	15,3%
	Assinatura Genômica	97,5%	13,5%

6.1.2 Efeito da diferença do conteúdo G+C entre doador e receptor

O impacto da similaridade entre o conteúdo G+C das ilhas genômicas e dos genomas receptores na predição das IGs foi estudado neste experimento. Os testes realizados apontaram taxas de 100% na identificação das ilhas inseridas de todas as combinações de subconjuntos (entre "A"e os demais), com exceção da combinação de "A"e "B", que resultou na identificação das ilhas em 9 de 10 sequências.

Os resultados alcançados no experimento mostram que o método é capaz de identificar com elevadas taxas de acerto ilhas genômicas provenientes de doadores com variações diversas de conteúdo G+C em relação ao receptor, ainda que essa diferença seja muito baixa.

6.1.3 Preenchimento automático dos fragmentos de fronteira

As taxas de acerto na identificação das ilhas genômicas presentes em fragmentos incompletos de 5 kb, 12,5 kb e 20 kb foram de 26%, 76% e 96%, respectivamente. Observa-se, portanto, que quanto maior for a proporção de nucleotídeos remanescentes em relação ao tamanho do fragmento (definido em 25 kb para o experimento), maior será a sensibilidade do método. Apesar da sua taxa de acerto na identificação de ilhas genômicas que compreendem 20% do tamanho do fragmento ter sido baixa, ainda assim considera-se a abordagem válida para esses casos, tendo em vista que o conteúdo da região de fronteira seria perdido caso não houvesse o seu preenchimento. No entanto, para regiões que possuem tamanho superior à metade do estabelecido para o fragmento, o método possui boa sensibilidade, alcançando até 96% de taxa de acerto. Em vista disso, considera-se a estratégia satisfatória para que ilhas genômicas presentes nos fragmentos incompletos da fronteira final do genoma não sejam ignoradas e possam ser identificadas.

6.1.4 Relação do tamanho dos fragmentos e das ilhas genômicas

Fragmentos de tamanho 25 kb e ilhas genômicas de três tamanhos diferentes foram utilizados no experimento definido na Seção 5.3.1, cujo resultado foi apresentado na Seção 6.1.1. A configuração do tamanho de janela permitiu que o método identificasse as ilhas genômicas com boas taxas de acerto, mesmo as de 10 kb (tamanho mínimo típico encontrado em IGs), que compreendem 30% do tamanho definido para o fragmento. O método também foi capaz de identificar ilhas genômicas de 50 kb, tamanho superior ao definido para o fragmento. Acredita-se, portanto, que o método é capaz de identificar com boa precisão, ilhas genômicas que compreendem pelo menos 30% do tamanho do fragmento, sendo a configuração de 25 kb uma opção viável, tendo em vista que a mesma tem boa eficiência tanto na identificação de ilhas menores, quanto na identificação das ilhas maiores.

6.2 Bactérias com IGs conhecidas

Os resultados obtidos nos testes realizados com cada uma das bactérias apresentadas na Tabela 5.1 são descritos nesta seção. O perfil cumulativo de G+C (curva z') [Zhang e Zhang

2004b] é utilizado para visualização dos resultados, pois essa abordagem é capaz de reproduzir características importantes de uma sequência genômica, cujo salto ou queda na curva indicam, respectivamente, aumento e diminuição abrupta no conteúdo G+C, apontando uma provável origem horizontal da região, embora também possa representar genes altamente expressos [de Brito et al. 2016]. Ademais, realiza-se uma breve descrição de cada uma das bactérias analisadas.

6.2.1 *Vibrio cholerae* chromosome II

Vibrio cholerae é o agente etiológico da cólera, que continua sendo uma séria ameaça a saúde das populações, principalmente nos países mais pobres, onde o saneamento básico é precário [Heidelberg et al. 2000]. O MSGIP identificou apenas uma IG para este genoma, a VCGI01, que possui 150 kb e parte da região do genoma iniciada em 0,300 Mb, prolongando-se até a posição 0,450 Mb. Os genes hospedados na região estão associados com cloranfenicol acetiltransferase, proteína *killer*, proteína antídoto, hemaglutinina e outras cópias de acetiltransferase. Observa-se um grande número de proteínas hipotéticas na região. Essa IG foi descoberta por [Nag et al. 2006]. O perfil cumulativo de G+C da bactéria é mostrado na Figura 6.1, com a IG identificada destacada em vermelho.

6.2.2 *Corynebacterium glutamicum* ATCC 13032

Corynebacterium glutamicum é uma bactéria que tem como principal característica a capacidade de produzir quantidades significativas de diferentes aminoácidos, cujo papel na indústria é fundamental [Kalinowski et al. 2003]. O tamanho do seu genoma é de aproximadamente 3,30 Mb. Quando analisada pelo MSGIP, cinco IGs foram identificadas.

A região CGGI01 corresponde às posições 0,625 - 0,650 Mb e contém uma grande quantidade de proteínas hipotéticas (cuja função é desconhecida), tipicamente encontradas em IGs [Langille et al. 2010] e elementos relatados em genomas de diferentes espécies, indicando uma possível origem horizontal da região. A IG CGGI02 abrange as posições de 0,750 - 0,775 Mb e contém duas proteínas hipotéticas e três helicases, dentre 7 elementos presentes na região. A CGGI03, iniciada na posição 1,775 - 2,000 Mb, corresponde a IG identificada previamente no trabalho de [Zhang e Zhang 2004b] e contém grande quantidade de proteínas

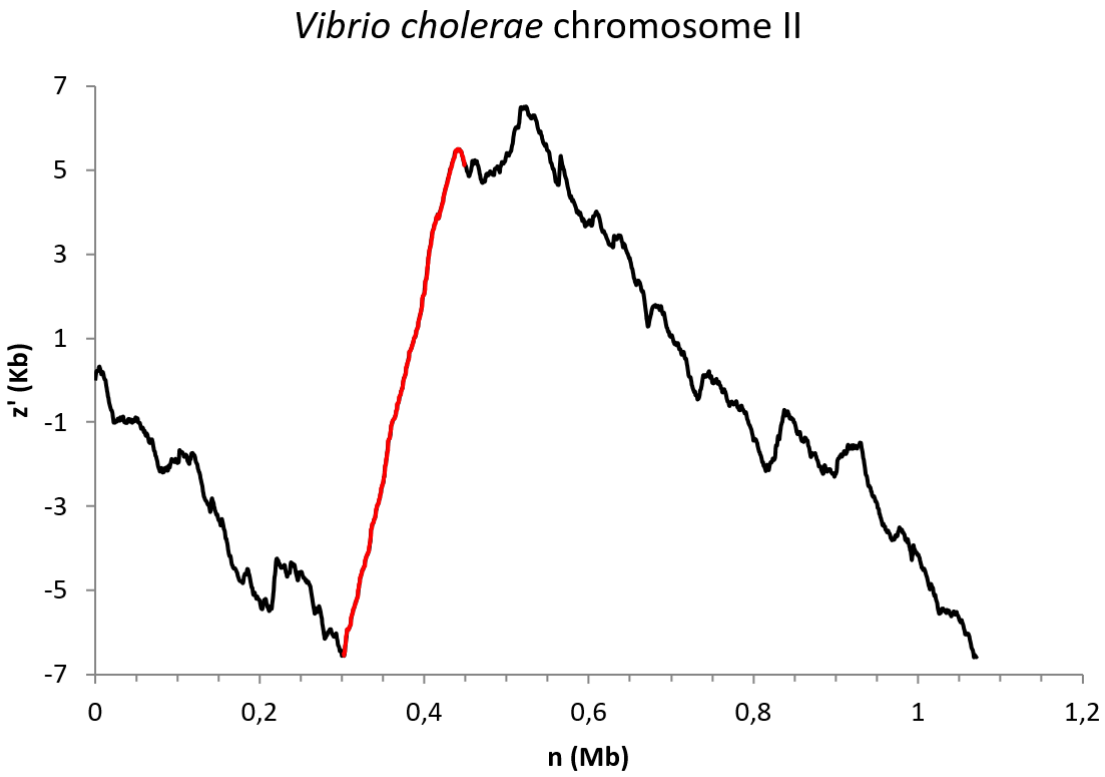

Figura 6.1: **Curva z' para o genoma da bactéria *Vibrio cholerae* chromosome II**

O eixo horizontal representa a variação do tamanho do genoma, em Mb, e o eixo vertical representa o valor da curva z', em kb, na respectiva posição do genoma. A linha preta corresponde a variação do perfil cumulativo de G+C ao longo do genoma e a linha em vermelho exibe as IGs identificadas pelo MSGIP.

Fonte: Elaborada pelo autor.

hipotéticas. Na região, 82,9% dos genes codificam este tipo de proteína, enquanto que no restante do genoma apenas 26,6% dos genes a codificam.

A IG identificada na sequência, 2,700 - 2,725 Mb (CGGI04), contém proteínas hipotéticas, elementos relatados em outras espécies, proteínas transportadoras e transferase. Essa ilha está sendo descrita pela primeira vez. Do mesmo modo, a última IG identificada, a CGGI05 (3,150 - 3,175 Mb) apresenta um elevado número de proteínas hipotéticas e elementos relatados em outras espécies, indicando a sua provável origem horizontal. A Figura 6.2 exibe a curva z' do genoma desta bactéria, cujas IGs descobertas são representadas por linhas vermelhas.

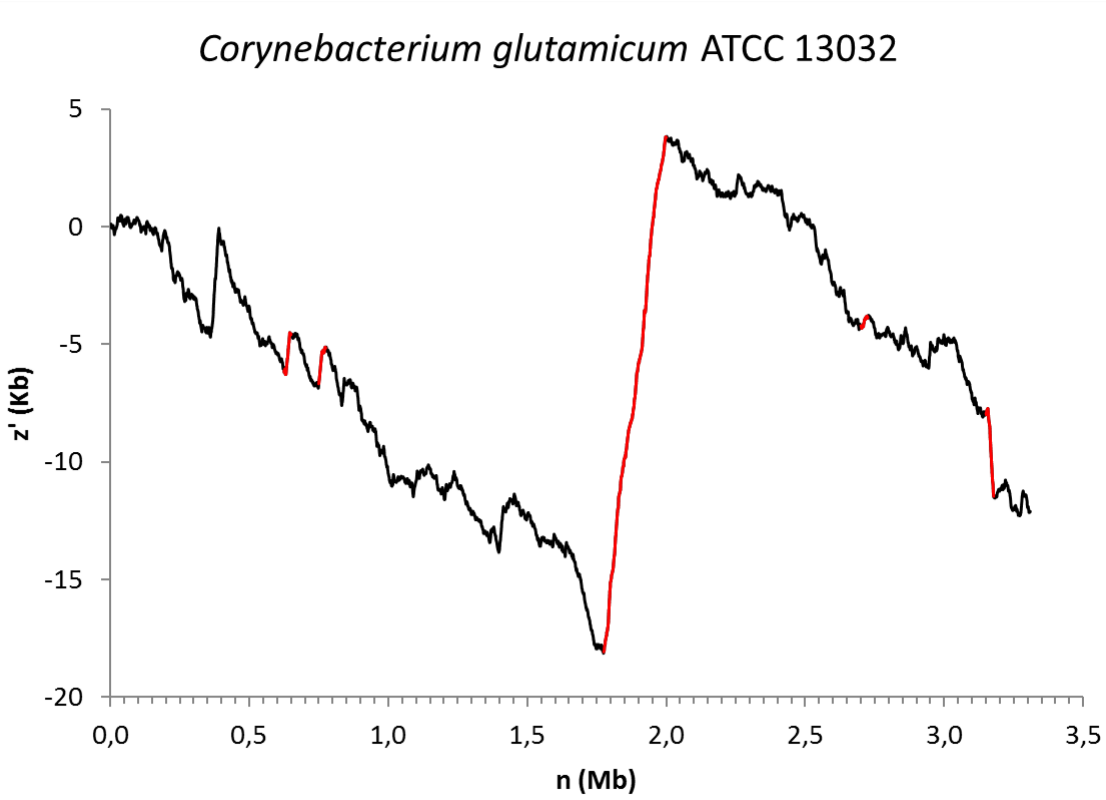

Figura 6.2: **Curva z' para o genoma da bactéria *Corynebacterium glutamicum* ATCC 13032**

O eixo horizontal representa a variação do tamanho do genoma, em Mb, e o eixo vertical representa o valor da curva z', em kb, na respectiva posição do genoma. A linha preta corresponde a variação do perfil cumulativo de G+C ao longo do genoma e a linha em vermelho exibe as IGs identificadas pelo MSGIP.

Fonte: Elaborada pelo autor.

6.2.3 *Streptococcus mutans* UA159

Streptococcus mutans é uma bactéria considerada a principal causa de cáries dentárias no mundo [Ajdić et al. 2002]. Seu genoma tem tamanho estimado em 2,03 Mb. A aplicação do MSGIP nesta sequência genômica retornou quatro regiões como prováveis IGs: SMGI01 (0,175 - 0,200 Mb), SMGI02 (1,250 - 1,300 Mb), SMGI03 (1,775 - 1,800 Mb) e SMGI04 (1,875 - 1,900 Mb). A partir da análise do conteúdo da primeira IG identificada (SMGI01), verifica-se a presença de integrase e transposase, elementos genéticos móveis frequentemente encontrados em ilhas genômicas [Che et al. 2014a], além da grande quantidade de

proteínas hipotéticas. A outra IG identificada (SMGI02) consiste da região *TnSmu2*, conhecida por ter um conteúdo G+C bem mais baixo do que o restante do genoma [Waterhouse e Russell 2006]. Essa região carrega genes que codificam peptídeos não ribossomais, síntese de policetídeos e proteínas acessórias responsáveis pela biossíntese do pigmento *mutanobactin* carregado pela *S. mutans*.

A IG SMGI03, encontrada na sequência, possui uma grande quantidade de proteínas hipotéticas, indicando uma provável origem horizontal, conforme apontado pelo método. Também contém uma bacteriocina, proteína de secreção que pode ser um fator de virulência, que é encontrado com maior frequência em IPs [Che et al. 2014a]. A última IG descoberta pelo método, a SMGI04, contém grande número de proteínas ribossomais, conhecidas por serem altamente expressas e que frequentemente levam a identificação de falsas IGs por métodos baseados em composição de sequência [Che et al. 2014a], como é o provável caso da região identificada pelo MSGIP. A Figura 6.3 mostra a representação gráfica do genoma da bactéria, destacando as IGs identificadas pelo MSGIP.

6.2.4 *Streptococcus pneumoniae*

Streptococcus pneumoniae é uma bactéria patogênica causadora de doenças graves, tais como sépsis, meningite e pneumonia [Bogaert et al. 2004]. Sua sequência genômica tem tamanho estimado de 2,08 Mb. Para este genoma, o MSGIP encontrou três regiões como potenciais IGs (SPGI01, SPGI02 e SPGI03). A primeira compreende a região 0,100 - 0,125 Mb e possui diversas proteínas hipotéticas. A segunda abrange as posições 0,925 - 0,950 Mb e contém um gene associado a fagos, reforçando a origem horizontal da região identificada [Che et al. 2014a]. Como discutido em [Guo e Wei 2012], essa região possui elementos que sugerem que ela seja uma ilha de patogenicidade, como proteínas com fatores relacionados ao sistema aquisição de ferro e protease de serina. A última IG identificada, que inicia em 1,225 Mb e termina em 1,250 Mb, possui uma grande quantidade de transposases, que são elementos típicos de regiões transferidas horizontalmente [Che et al. 2014a]. Tanto a região SPGI02, quanto a SPGI03 foram discutidas no trabalho de [Guo e Wei 2012]. A representação gráfica (curva z') da bactéria é apresentadas na Figura 6.4, em que as IGs descobertas são realçadas por linhas vermelhas.

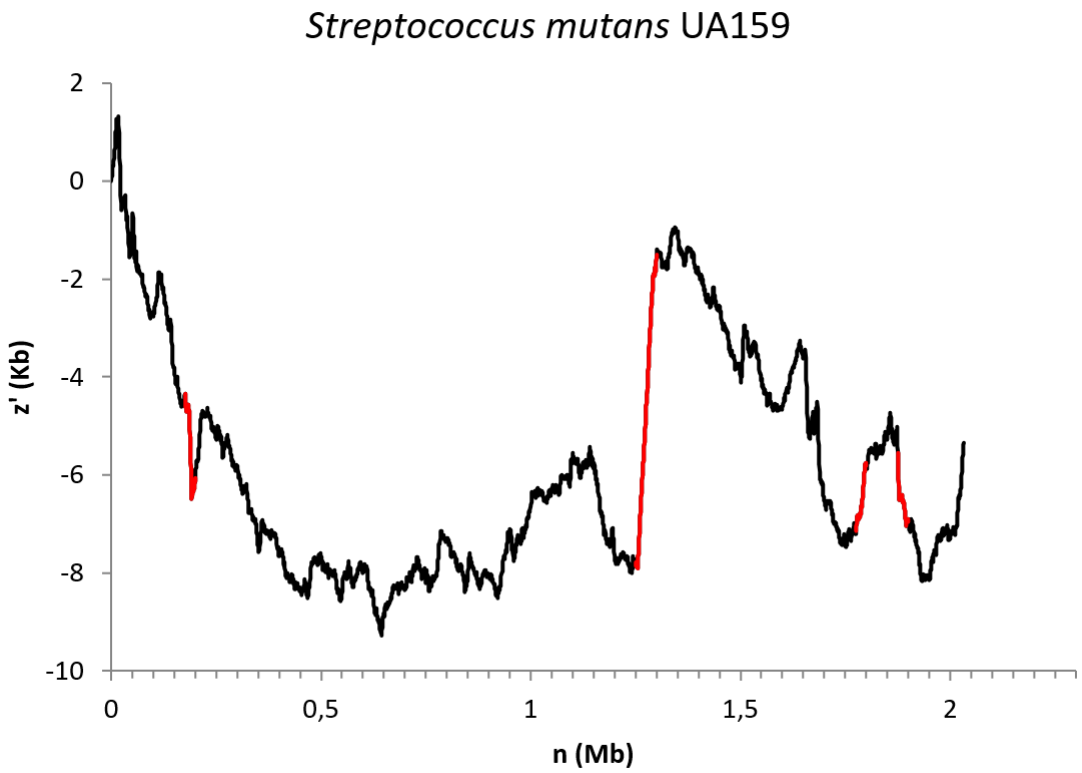

Figura 6.3: **Curva z' para o genoma da bactéria *Streptococcus mutans* UA159**

O eixo horizontal representa a variação do tamanho do genoma, em Mb, e o eixo vertical representa o valor da curva z', em kb, na respectiva posição do genoma. A linha preta corresponde a variação do perfil cumulativo de G+C ao longo do genoma e a linha em vermelho exibe as IGs identificadas pelo MSGIP.

Fonte: Elaborada pelo autor.

6.2.5 *Rhodopseudomonas palustris* CGA009

Rhodopseudomonas palustris é uma bactéria com tamanho aproximado de 5,46 Mb, conhecida pela sua versatilidade metabólica, sendo capaz de prover diferentes tipos de metabolismo. Dispõe de aplicações importantes como biodegradação e produção de hidrogênio [Larimer et al. 2004]. Para esta bactéria, o MSGIP encontrou seis IGs, sendo duas delas previamente descritas na literatura. A primeira região identificada (1,475 - 1,500 Mb), denominada RPGI01, apresenta um grande número de proteínas hipotéticas, corroborando o indício de sua origem em outros organismos. A RGPI02 (2,125 - 2,150 Mb) possui várias proteínas hipotéticas e genes relacionados a fagos, evidenciando a sua putativa origem

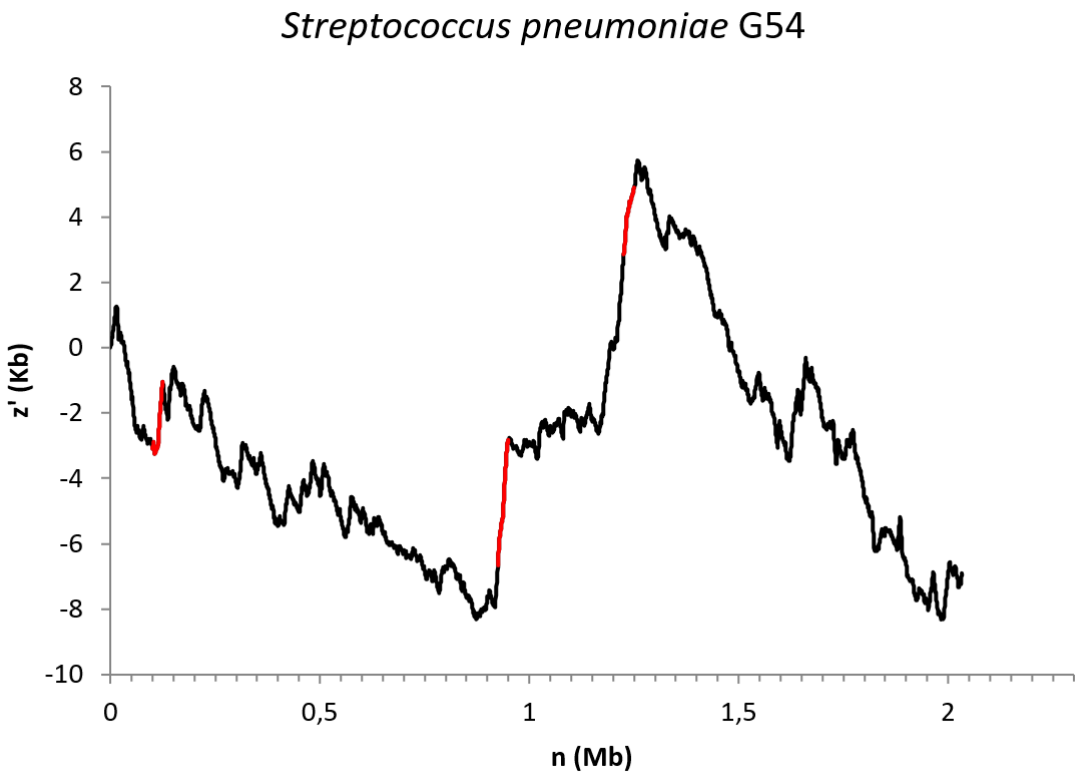

Figura 6.4: **Curva z' para o genoma da bactéria *Streptococcus pneumoniae***

O eixo horizontal representa a variação do tamanho do genoma, em Mb, e o eixo vertical representa o valor da curva z', em kb, na respectiva posição do genoma. A linha preta corresponde a variação do perfil cumulativo de G+C ao longo do genoma e a linha em vermelho exibe as IGs identificadas pelo MSGIP.

Fonte: Elaborada pelo autor.

horizontal. A RPGI03 (2,700 - 2,725 Mb) apresenta abundância de proteínas hipotéticas e proteínas transportadoras. As primeiras três IGs estão sendo descritas pela primeira vez.

Na sequência, a RPGI04 (3,650 - 3,700 Mb) incorpora grande número de proteínas ribossomais conhecidas por serem altamente expressas e que, frequentemente, levam a identificação de falsas IGs por métodos baseados em composição de sequência [Che et al. 2014a]. Apesar da região ter sido identificada pelo MSGIP, acredita-se que a mesma não possui origem horizontal, sendo caracterizada como um falso positivo. A IG RPGI05 (3,750 - 3,800 Mb) hospeda genes associados a proteínas hipotéticas e consiste de uma ilha genômica previamente descrita [Zhang e Zhang 2004a]. Por último, o método identificou a IG RPGI06

(4,575 - 4,625 Mb), que hospeda uma grande quantidade de proteínas hipotéticas e proteínas transportadoras ABC, evidenciando a sua provável origem horizontal. Essa IG foi igualmente encontrada no trabalho de [Zhang e Zhang 2004a]. As IGs identificadas podem ser visualizadas na Figura 6.5.

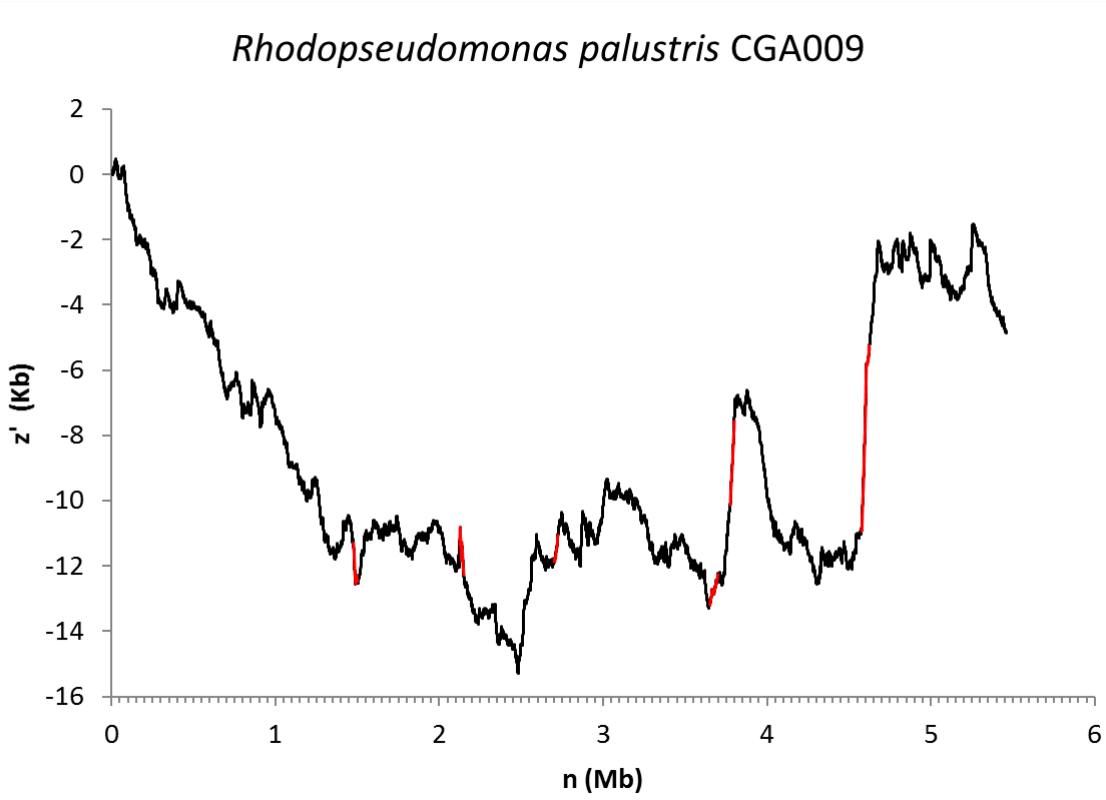

Figura 6.5: **Curva z' para o genoma da bactéria *Rhodopseudomonas palustris* CGA009**

O eixo horizontal representa a variação do tamanho do genoma, em Mb, e o eixo vertical representa o valor da curva z', em kb, na respectiva posição do genoma. A linha preta corresponde a variação do perfil cumulativo de G+C ao longo do genoma e a linha em vermelho exibe as IGs identificadas pelo MSGIP.

Fonte: Elaborada pelo autor.

Este capítulo apresentou os resultados obtidos a partir da execução do MSGIP nas duas categorias de experimentos propostos e descritos no capítulo anterior. A análise da sua execução com sequências quiméricas com ilhas genômicas embutidas evidenciou a sua superioridade em relação aos métodos tradicionais de predição, enquanto que sua aplicação em genomas com IGs conhecidas indicou que o mesmo é capaz de produzir resultados em conformidade com os apresentados em outros trabalhos da área, além de propiciar a identificação de regiões ainda não conhecidas.

Capítulo 7

Considerações Finais

Este trabalho apresentou a proposta de um novo método para predição de ilhas genômicas em bactérias, fundamentado no algoritmo de agrupamento *mean shift*. O método requer que o usuário defina apenas o tamanho dos fragmentos a serem utilizados como unidade básica de predição e a unidade de composição de sequência utilizada no processo de identificação das regiões, que são palavras de tamanho k, cujo alfabeto é constituído pelas bases nucléicas (A, T, C, G).

Com vistas a avaliar a capacidade do método na predição de IGs em diferentes situações, dois experimentos foram propostos. Os resultados obtidos em cada um deles atestam a capacidade do método em identificar as ilhas genômicas. O primeiro, realizado com sequências quiméricas com ilhas genômicas embutidas, comprovou a capacidade do método na identificação de regiões do genoma composicionalmente distintas em vários cenários práticos possíveis. Finalmente, os testes realizados com bactérias com IGs conhecidas mostram que o método produz resultados consistentes com os descritos na literatura, demonstrando a sua eficácia e aplicabilidade na identificação de ilhas genômicas em bactérias. O mesmo igualmente foi capaz de identificar novas regiões, não encontradas por outros métodos existentes na literatura, cuja origem horizontal foi hipotetizada com base na análise dos genes presentes nas regiões.

Espera-se que a abordagem apresentada neste trabalho seja difundida e que auxilie pesquisadores na análise de ilhas genômicas em diversos organismos. Acredita-se que ela possa ser utilizada como base no desenvolvimento de novos métodos de predição IGs e em combinação com outros métodos já existentes.

7.1 Trabalhos Futuros

Como propostas de trabalhos futuros, citam-se:

- Desenvolver e disponibilizar uma interface web intuitiva para que o método apresentado possa ser acessado facilmente por pesquisadores, sem necessidade de *download* e instalação de nenhum tipo de aplicação.
- Estudar meios de refinar as fronteiras das ilhas genômicas encontradas em uma dada sequência genômica.
- Analisar a estratégia utilizada no cálculo do parâmetro *bandwidth* para os casos específicos de análise de bactérias com genomas pequenos.
- Estudar com maior profundidade o impacto da variação do tamanho dos fragmentos e definir automaticamente os parâmetros do método de acordo com o valor selecionado.

Bibliografia

[Ajdić et al. 2002] Ajdić, D., McShan, W. M., McLaughlin, R. E., Savić, G., Chang, J., Carson, M. B., Primeaux, C., Tian, R., Kenton, S., Jia, H., et al. (2002). Genome sequence of streptococcus mutans ua159, a cariogenic dental pathogen. *Proceedings of the National Academy of Sciences*, 99(22):14434–14439.

[Alberts et al. 2015] Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., Watson, J. D., e Grimstone, A. (2015). Molecular biology of the cell (6rd edn).

[Andersson 2009] Andersson, J. O. (2009). Gene transfer and diversification of microbial eukaryotes. *Annual review of microbiology*, 63:177–193.

[Attwood et al. 2011] Attwood, T., Gisel, A., Bongcam-Rudloff, E., e Eriksson, N. (2011). *Concepts, historical milestones and the central place of bioinformatics in modern biology: a European perspective*. INTECH Open Access Publisher.

[Basu et al. 2004] Basu, S., Bilenko, M., e Mooney, R. J. (2004). A probabilistic framework for semi-supervised clustering. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 59–68. ACM.

[Bennett et al. 2016] Bennett, G. M., Abbà, S., Kube, M., e Marzachì, C. (2016). Complete genome sequences of the obligate symbionts "candidatus sulcia muelleri" and "ca. nasuia deltocephalinicola" from the pestiferous leafhopper macrosteles quadripunctulatus (hemiptera: Cicadellidae). *Genome announcements*, 4(1):e01604–15.

[Berkhin 2006a] Berkhin, P. (2006a). A survey of clustering data mining techniques. In *Grouping multidimensional data*, pages 25–71. Springer.

[Berkhin 2006b] Berkhin, P. (2006b). A survey of clustering data mining techniques. In *Grouping multidimensional data*, pages 25–71. Springer.

[Bogaert et al. 2004] Bogaert, D., de Groot, R., e Hermans, P. (2004). Streptococcus pneumoniae colonisation: the key to pneumococcal disease. *The Lancet infectious diseases*, 4(3):144–154.

[Che et al. 2014a] Che, D., Hasan, M., e Chen, B. (2014a). Identifying Pathogenicity Islands in Bacterial Pathogenomics Using Computational Approaches. *Pathogens*, 3(1):36–56.

[Che et al. 2011] Che, D., Hasan, M. S., Wang, H., Fazekas, J., Huang, J., e Liu, Q. (2011). Egid: an ensemble algorithm for improved genomic island detection in genomic sequences. *Bioinformation*, 7(6):311.

[Che et al. 2014b] Che, D., Wang, H., Fazekas, J., e Chen, B. (2014b). An accurate genomic island prediction method for sequenced bacterial and archaeal genomes. *Journal of Proteomics & Bioinformatics*, 7(8):214.

[Cheng 1995] Cheng, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):790–799.

[Cristianini e Hahn 2006] Cristianini, N. e Hahn, M. W. (2006). *Introduction to computational genomics: a case studies approach*. Cambridge University Press.

[Das et al. 2008] Das, S., Abraham, a., e Konar, a. (2008). Automatic Clustering Using an Improved Differential Evolution Algorithm. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 38(1):218–237.

[de Brito et al. 2016] de Brito, D. M., Maracaja-Coutinho, V., de Farias, S. T., Batista, L. V., e do Rêgo, T. G. (2016). A novel method to predict genomic islands based on mean shift clustering algorithm. *PloS one*, 11(1):e0146352.

[Dhillon et al. 2015] Dhillon, B. K., Laird, M. R., Shay, J. a., Winsor, G. L., Lo, R., Nizam, F., Pereira, S. K., Waglechner, N., McArthur, A. G., Langille, M. G., e Brinkman, F. S. (2015). IslandViewer 3: more flexible, interactive genomic island discovery, visualization and analysis: Figure 1. *Nucleic Acids Research*, 43(W1):W104–W108.

[Doolittle 1999] Doolittle, W. F. (1999). Phylogenetic classification and the universal tree. *Science (New York, N.Y.)*, 284(5423):2124–2129.

[Faceli et al. 2011] Faceli, K., Lorena, C. A., Gama, J., e Carvalho, André, C. P. L. F. (2011). *Inteligência Artificial: Uma abordagem de aprendizado de máquina*. Grupo Gen-LTC.

[Fleischmann et al. 1995] Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J.-F., Dougherty, B. A., Merrick, J. M., et al. (1995). Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science*, 269(5223):496–512.

[Frigui et al. 1999] Frigui, H., Krishnapuram, R., e Member, S. (1999). With Applications in Computer Vision. 21(5):450–465.

[Georgescu et al. 2003] Georgescu, B., Shimshoni, I., e Meer, P. (2003). Mean Shift Based Clustering in High Dimensions: A Texture Classification Example. *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 456–463.

[Gladieux et al. 2014] Gladieux, P., Ropars, J., Badouin, H., Branca, A., Aguileta, G., Vienne, D. M., Rodríguez de la Vega, R. C., Branco, S., e Giraud, T. (2014). Fungal evolutionary genomics provides insight into the mechanisms of adaptive divergence in eukaryotes. *Molecular ecology*, 23(4):753–773.

[Guo e Wei 2012] Guo, F.-B. e Wei, W. (2012). Prediction of genomic islands in three bacterial pathogens of pneumonia. *International journal of molecular sciences*, 13(3):3134–3144.

[Guzella e Caminhas 2009] Guzella, T. S. e Caminhas, W. M. (2009). A review of machine learning approaches to spam filtering. *Expert Systems with Applications*, 36(7):10206–10222.

[Hacker e Kaper 2000] Hacker, J. e Kaper, J. B. (2000). Pathogenicity islands and the evolution of microbes. *Annual Reviews in Microbiology*, 54(1):641–679.

[Han et al. 2011] Han, J., Kamber, M., e Pei, J. (2011). *Data mining: concepts and techniques*. Elsevier.

[Hasan et al. 2012] Hasan, M. S., Liu, Q., Wang, H., Fazekas, J., Chen, B., e Che, D. (2012). Gist: Genomic island suite of tools for predicting genomic islands in genomic sequences. *Bioinformation*, 8(4):203.

[Heidelberg et al. 2000] Heidelberg, J. F., Eisen, J. A., Nelson, W. C., Clayton, R. A., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickey, E. K., Peterson, J. D., Umayam, L., et al. (2000). Dna sequence of both chromosomes of the cholera pathogen vibrio cholerae. *Nature*, 406(6795):477–483.

[Hentschel e Hacker 2001] Hentschel, U. e Hacker, J. (2001). Pathogenicity islands: the tip of the iceberg. *Microbes and infection*, 3(7):545–548.

[Hsiao et al. 2003] Hsiao, W., Wan, I., Jones, S. J., e Brinkman, F. S. (2003). Islandpath: aiding detection of genomic islands in prokaryotes. *Bioinformatics*, 19(3):418–420.

[Jain e Dubes 1988] Jain, A. K. e Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc.

[Jain et al. 1999] Jain, A. K., Murty, M. N., e Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323.

[Juhas et al. 2009] Juhas, M., van der Meer, J. R., Gaillard, M., Harding, R. M., Hood, D. W., e Crook, D. W. (2009). Genomic islands: tools of bacterial horizontal gene transfer and evolution. *FEMS microbiology reviews*, 33(2):376–393.

[Jung et al. 2003] Jung, Y., Park, H., Du, D.-Z., e Drake, B. L. (2003). A decision criterion for the optimal number of clusters in hierarchical clustering. *Journal of Global Optimization*, 25(1):91–111.

[Kalinowski et al. 2003] Kalinowski, J., Bathe, B., Bartels, D., Bischoff, N., Bott, M., Burkovski, A., Dusch, N., Eggeling, L., Eikmanns, B. J., Gaigalat, L., et al. (2003). The complete corynebacterium glutamicum atcc 13032 genome sequence and its impact on the production of l-aspartate-derived amino acids and vitamins. *Journal of biotechnology*, 104(1):5–25.

[Karlin 2001] Karlin, S. (2001). Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends in microbiology*, 9(7):335–343.

[Keeling e Palmer 2008] Keeling, P. J. e Palmer, J. D. (2008). Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics*, 9(8):605–618.

[Langille et al. 2008] Langille, M. G., Hsiao, W. W., e Brinkman, F. S. (2008). Evaluation of genomic island predictors using a comparative genomics approach. *BMC bioinformatics*, 9(1):329.

[Langille et al. 2010] Langille, M. G., Hsiao, W. W., e Brinkman, F. S. (2010). Detecting genomic islands using bioinformatics approaches. *Nature Reviews Microbiology*, 8(5):373–382.

[Larimer et al. 2004] Larimer, F. W., Chain, P., Hauser, L., Lamerdin, J., Malfatti, S., Do, L., Land, M. L., Pelletier, D. A., Beatty, J. T., Lang, A. S., et al. (2004). Complete genome sequence of the metabolically versatile photosynthetic bacterium *Rhodopseudomonas palustris*. *Nature biotechnology*, 22(1):55–61.

[Larranaga et al. 2006] Larranaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J. A., Armañanzas, R., Santafé, G., Pérez, A., et al. (2006). Machine learning in bioinformatics. *Briefings in bioinformatics*, 7(1):86–112.

[Lawrence e Ochman 1998] Lawrence, J. G. e Ochman, H. (1998). Molecular archaeology of the *Escherichia coli* genome. *Proceedings of the National Academy of Sciences*, 95(16):9413–9417.

[Lawrence e Roth 1996] Lawrence, J. G. e Roth, J. R. (1996). Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics*, 143(4):1843–1860.

[Lewin 2004] Lewin, B. (2004). *genes VIII*. Pearson Prentice Hall.

[Lightfield et al. 2011] Lightfield, J., Fram, N. R., e Ely, B. (2011). Across bacterial phyla, distantly-related genomes with similar genomic gc content have similar patterns of amino acid usage. *PloS one*, 6(3):e17677.

[Madigan et al. 2015] Madigan, M. T., Martinko, J. M., Bender, K. S., Buckley, D. H., e Stahl, D. A. (2015). Brock biology of microorganisms, 14th edn.

[Marcet-Houben e Gabaldón 2010] Marcet-Houben, M. e Gabaldón, T. (2010). Acquisition of prokaryotic genes by fungal genomes. *Trends in Genetics*, 26(1):5–8.

[Merceron e Yacef 2005] Merceron, A. e Yacef, K. (2005). Clustering students to help evaluate learning. In *Technology Enhanced Learning*, pages 31–42. Springer.

[Mitchell et al. 1997] Mitchell, T. M. et al. (1997). Machine learning.

[Mitić et al. 2008] Mitić, N. S., Pavlović-Lažetić, G. M., e Beljanski, M. V. (2008). Could n-gram analysis contribute to genomic island determination? *Journal of Biomedical Informatics*, 41(6):936–943.

[Nag et al. 2006] Nag, S., Chatterjee, R., Chaudhuri, K., e Chaudhuri, P. (2006). Unsupervised statistical identification of genomic islands using oligonucleotide distributions with application to*Vibrio* genomes. *Sadhana*, 31(2):105–115.

[Nelson et al. 1999] Nelson, K. E., Clayton, R. A., Gill, S. R., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickey, E. K., Peterson, J. D., Nelson, W. C., Ketchum, K. A., et al. (1999). Evidence for lateral gene transfer between archaea and bacteria from genome sequence of thermotoga maritima. *Nature*, 399(6734):323–329.

[Organização Mundial de Saúde 2017] Organização Mundial de Saúde (2017). Who publishes list of bacteria for which new antibiotics are urgently needed. http://www.who.int/mediacentre/news/releases/2017/bacteria-antibiotics-needed/en/m. Acesso em: 21 mar 2017.

[Ou et al. 2007] Ou, H.-Y., He, X., Harrison, E. M., Kulasekara, B. R., Thani, A. B., Kadioglu, A., Lory, S., Hinton, J. C., Barer, M. R., Deng, Z., et al. (2007). Mobilomefinder: web-based tools for in silico and experimental discovery of bacterial genomic islands. *Nucleic acids research*, 35(suppl 2):W97–W104.

[Rajan et al. 2007] Rajan, I., Aravamuthan, S., e Mande, S. S. (2007). Identification of compositionally distinct regions in genomes using the centroid method. *Bioinformatics*, 23(20):2672–2677.

[Russell et al. 2011] Russell, S., Norvig, P., e Intelligence, A. (2011). Artificial intelligence: A modern approach 3.ed.

[Sasaki et al. 2014] Sasaki, H., Hyvärinen, a., e Sugiyama, M. (2014). Clustering via Mode Seeking by Direct Estimation of the Gradient of a Log-Density. *arXiv preprint arXiv:1404.5028*, pages 1–18.

[Shrivastava et al. 2010] Shrivastava, S., Siva Kumar Reddy, C. V., e Mande, S. S. (2010). INDeGenIUS, a new method for high-throughput identification of specialized functional islands in completely sequenced organisms. *Journal of Biosciences*, 35(3):351–364.

[Sisodia et al. 2012] Sisodia, D., Singh, L., Sisodia, S., e Saxena, K. (2012). Clustering techniques: A brief survey of different clustering algorithms. *International Journal of Latest Trends in Engineering and Technology (IJLTET)*, 1(3):82–87.

[Soares et al. 2012] Soares, S. C., Abreu, V. a. C., Ramos, R. T. J., Cerdeira, L., Silva, A., Baumbach, J., Trost, E., Tauch, A., Hirata, R., Mattos-Guaraldi, A. L., Miyoshi, A., e Azevedo, V. (2012). PIPS: Pathogenicity Island Prediction Software. *PLoS ONE*, 7(2):e30848.

[Soares et al. 2015] Soares, S. C., Geyik, H., Ramos, R. T., de Sá, P. H., Barbosa, E. G., Baumbach, J., Figueiredo, H. C., Miyoshi, A., Tauch, A., Silva, A., e Azevedo, V. (2015). GIPSy: Genomic island prediction software. *Journal of Biotechnology*.

[Tsai et al. 2014] Tsai, C.-W., Lin, T.-H., e Chiang, M.-C. (2014). Automatic elastic net clustering algorithm. In *Systems, Man and Cybernetics (SMC), 2014 IEEE International Conference on*, pages 2768–2773. IEEE.

[Tsai et al. 2013] Tsai, C.-W., Tai, C.-A., e Chiang, M.-C. (2013). An automatic data clustering algorithm based on differential evolution. In *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*, pages 794–799. IEEE.

[Tu e Ding 2003] Tu, Q. e Ding, D. (2003). Detecting pathogenicity islands and anomalous gene clusters by iterative discriminant analysis. *FEMS microbiology letters*, 221(2):269–275.

[van Dijk et al. 2014] van Dijk, E. L., Auger, H., Jaszczyszyn, Y., e Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in genetics*, 30(9):418–426.

[Venter et al. 2001] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., et al. (2001). The sequence of the human genome. *science*, 291(5507):1304–1351.

[Waack et al. 2006] Waack, S., Keller, O., Asper, R., Brodag, T., Damm, C., Fricke, W. F., Surovcik, K., Meinicke, P., e Merkl, R. (2006). Score-based prediction of genomic islands in prokaryotic genomes using hidden markov models. *BMC bioinformatics*, 7(1):142.

[Waterhouse e Russell 2006] Waterhouse, J. C. e Russell, R. R. (2006). Dispensable genes and foreign dna in *Streptococcus mutans*. *Microbiology*, 152(6):1777–1788.

[Wu e Yang 2007] Wu, K.-L. e Yang, M.-S. (2007). Mean shift-based clustering. *Pattern Recognition*, 40(11):3035–3052.

[Zhang e Zhang 2004a] Zhang, C.-T. e Zhang, R. (2004a). Genomic islands in *Rhodopseudomonas palustris*. *Nature biotechnology*, 22(9):1078–1079.

[Zhang e Zhang 2004b] Zhang, R. e Zhang, C.-T. (2004b). A systematic method to identify genomic islands and its applications in analyzing the genomes of *Corynebacterium glutamicum* and *Vibrio vulnificus* cmcp6 chromosome i. *Bioinformatics*, 20(5):612–622.