Detecção de ilhas genômicas em procariotos utilizando o método *Expectation-Maximization*

Tayrone de Sousa Monteiro

CENTRO DE INFORMÁTICA
UNIVERSIDADE FEDERAL DA PARAÍBA

João Pessoa, PB
Dezembro - 2017

Tayrone de Sousa Monteiro

Detecção de ilhas genômicas em procariotos utilizando o método *Expectation-Maximization*

Monografia apresentada ao curso de Engenharia da Computação do Centro de Informática, da Universidade Federal da Paraíba, como requisito para a obtenção do grau de Bacharel em Engenheiro de Computação.

Orientador: Thaís Gaudencio do Rêgo

João Pessoa, PB
Dezembro - 2017

Catalogação na publicação
Seção de Catalogação e Classificação

M775d Monteiro, Tayrone de Sousa.
Detecção de ilhas genômicas em procariotos utilizando o método Expectation-Maximization / Tayrone de Sousa Monteiro. - João Pessoa, 2018.
34 f. : il.

Orientação: Thais Gaudêncio do Rêgo.
Monografia (Graduação) - UFPB/CI.

1. expectation-maximization, ilhas genômicas, cluster. I. do Rêgo, Thais Gaudêncio. II. Título.

UFPB/BC

CENTRO DE INFORMÁTICA
UNIVERSIDADE FEDERAL DA PARAÍBA

Trabalho de Conclusão de Curso de Tayrone de Sousa Monteiro, intitulado de **Detecção de ilhas genômicas em procariotos utilizando o método *Expectation-Maximization*,** de autoria de **Tayrone de Sousa Monteiro** aprovada pela banca examinadora constituída pelos seguintes professores:

_____

Prof. Thaís Gaudencio do Rêgo
Centro de Informática, UFPB

_____

Prof. Daniel Miranda de Brito
Centro de Informática, UFPB

_____

Prof. Lincoln David Nery e Silva
Centro de Ciências Exatas e da Natureza, UFPB

João Pessoa, 1 de dezembro de 2017

Centro de Informática, Universidade Federal da Paraíba
Rua dos Escoteiros, Mangabeira VII, João Pessoa, Paraíba, Brasil CEP: 58058-600
Fone: +55 (83) 3216 7093 / Fax: +55 (83) 3216 7117

RESUMO

A pesquisa aqui apresentada tem como objetivo utilizar um método de clusterização para a identificação de ilhas genômicas em procariotos. Dessa forma, se baseia em resultados obtidos por outros autores utilizando outras metodologias já conhecidas na literatura, objetivando reproduzir estes resultados com o método de clusterização Expectation-Maximization (EM). Para isso, o método foi implementado em linguagem Java, com auxílio de bibliotecas disponíveis pelo software Weka. Os testes foram realizados utilizando o mesmo conjunto de dados usados por Brito et al. (2016). Os resultados obtidos foram comparados com os de outros métodos presentes na literatura, comprovando a eficácia do EM. Por fim, a análise de algumas ilhas identificadas pelo método proposto, que não foram documentadas anteriormente, foi proposta como trabalho futuro.

Palavras-chave: Expectation-maximization, ilhas genômicas, cluster.

ABSTRACT

This research aims to utilize a clustering method for identifying genomic islands in prokaryotes. Based on results already obtained by many authors, the Expectation-Maximization method is applied in order to acquire similar results. The method was implemented in Java language, using libraries available by Weka software. All tests were done considering the same dataset used by Brito et al. (2016). The efficiency of the method could be proved by comparing its results to the others available on the literature. The analysis of some genomic islands that were not identified by the method is suggested as further work.

Key-words: Expectation-maximization, genomic islands, cluster.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE EQUAÇÕES

LISTA DE ABREVIATURAS

IGs	–	Ilhas Genômicas
TGH	–	Transferência genômica horizontal

SUMÁRIO

1. INTRODUÇÃO

Os procariotos são seres unicelulares de dimensões microscópicas que diferem dos organismos eucariotos basicamente por não apresentarem material genético delimitado por uma membrana. Além disso, não possuem boa parte das organelas presentes nos eucariotos, como as mitocôndrias e o Complexo de Golgi, por exemplo (KERFELD, 2005). Estes seres são classificados em dois domínios diferentes: Bacteria e Archaea. Com o desenvolvimento da tecnologia nas últimas décadas, e principalmente da bioinformática, foi possível realizar o sequenciamento genético de seres vivos e, naturalmente, de seres unicelulares, como as bactérias. Desde então, o número de espécies de bactérias que tiveram seu DNA sequenciado cresceu vertiginosamente nos últimos anos. Isto permitiu um melhor estudo do material genético desses organismos, resultando na descoberta de regiões de DNA chamadas de ilhas genômicas (MADIGAN et al., 2015).

Ilhas genômicas (IGs) são regiões de um genoma que foram adquiridas através de um processo de transferência horizontal. Ou seja, essas regiões foram transmitidas de um organismo para outro que não é seu descendente. Em bactérias, as IGs muitas vezes correspondem a materiais genéticos pertencentes originalmente a outras espécies. Nesses casos, o processo de transferência pode acarretar na assimilação de novas características e funções por parte do receptor do material transferido. Este fenômeno pode apresentar um resultado positivo para os organismos afetados, pois aumenta o fator de diversificação destes organismos e, consequentemente, contribue para a adaptação da espécie ao meio em que vive (JUHAS et al., 2009).

Este fator de adaptação é de grande interesse da comunidade científica, especialmente para a indústria farmacêutica, tendo em vista que resistência a certos antibióticos pode ser algo desenvolvido por bactérias através do processo de transferência genômica horizontal (TGH). Além disso, existem outros casos possíveis, como a possibilidade de obtenção de um material genético que irá desenvolver um caráter patogênico no organismo receptor, por exemplo (HACKER; KAPER, 2000).

Dessa forma, podemos afirmar que, para minimizar possíveis problemas que venham ser causados à saúde humana através da transferência genômica horizontal em bactérias, é necessário que a comunidade científica saiba identificar ilhas genômicas em um DNA, pois este é o primeiro passo para anular possíveis consequências negativas causadas pela presença de um trecho específico de genoma em uma certa espécie de bactéria.

Com a consciência de que a identificação de IGs é uma atividade que não pode ser ignorada na bioinformática, sabemos que alguns métodos já foram desenvolvidos com o objetivo de suprir esta necessidade. Tendo em vista as estratégias já desenvolvidas até o momento e suas limitações, este trabalho pretende apresentar uma alternativa a estas estratégias. Assim, foi desenvolvida uma técnica de identificação baseada no algoritmo de clusterização *Expectation-Maximization* (EM). No decorrer do trabalho, os resultados obtidos pelo EM são confrontados com os resultados de outras abordagens presentes na literatura, para que sua eficácia possa ser comprovada.

1.1 OBJETIVOS

Objetivo Geral: Aplicação do algoritmo de clusterização *Expectation-Maximization* (EM) como solução para o problema de identificação de ilhas genômicas em procariotos.

Objetivos Específicos:

- Realizar implementação do EM em uma linguagem de alto nível;
- Dados os resultados da clusterização do EM, definir uma regra para a diferenciação dos *clusters* que englobam IGs dentre aqueles que não contém IGs;
- Identificadas as ilhas genômicas, comparar os resultados obtidos com resultados de outros métodos da literatura.

Este trabalho está organizado nas seguintes seções: conceitos gerais e revisão de literatura; metodologia; apresentação e análise de resultados e, por fim, conclusão e trabalhos futuros.

2. CONCEITOS GERAIS E REVISÃO DA LITERATURA

O DNA pode ser definido como uma molécula que carrega informações genéticas sobre o desenvolvimento, funcionamento e reprodução de um organismo vivo (além de alguns tipos de vírus). Essa molécula (Figura 1) apresenta uma estrutura de dupla-hélice, apresentando duas fitas que se entrelaçam. Cada fita é composta por uma sequência de nucleotídeos, onde cada nucleotídeo pode ser composto por uma base nitrogenada dentre as quatro seguintes: adenina, guanina, citosina e timina. Os nucleotídeos se ligam um ao outro através de ligações covalentes, formando uma fita de DNA. Além disso, as bases de fitas diferentes também se ligam umas às outras, formando sempre as seguintes duplas: adenina e timina, citosina e guanina (ALBERTS et al., 2014). Essas ligações fazem com que a estrutura final do DNA seja a de duas fitas ligadas entre si, como é possível ver na imagem abaixo.

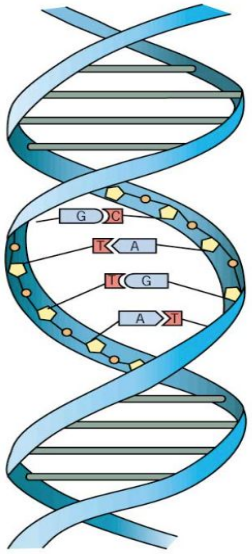

Figura 1. Estrutura do DNA.

Fonte: SILVA et al., 2015.

O primeiro sequenciamento completo de um genoma baseado em DNA foi realizado em 1977. O organismo em questão foi o vírus phiX174, que apresenta seu material genético disposto em uma única fita de DNA (SANGER et al., 1997). A partir de então, foi possível realizar grandes avanços no estudo de genomas de outros organismos, incluindo as bactérias. Em 1995, o primeiro organismo de vida livre teve seu DNA sequenciado pela equipe do microbiologista Hamilton Smith, que utilizou um exemplar da bactéria *Haemophilus influenzae* (FLEISCHMANN et al., 1995).

Hoje em dia, contamos com a disponibilidade de milhares de organismos sequenciados em diversas bases de dados na internet. O Instituto Nacional de Saúde (*National Institute of Health*, NIH) é principal órgão americano responsável pelo desenvolvimento de pesquisa em medicina e saúde pública nos Estados Unidos. Uma das suas ramificações é o Centro Nacional de Informação em Biotecnologia (*National Center for Biotechnology Information*, NCBI), setor responsável por disponibilizar um conjunto de bases de dados relacionadas à biotecnologia e biomedicina, além de prover serviços na área de bioinformática, que têm grande importância para a comunidade científica nos dias de hoje. Atualmente, o NCBI dispõe um conjunto de mais de 100.000 genomas de procariotos sequenciados em suas bases de dados (WANG; BRYANT, 2014). Assim, podemos afirmar que a disponibilidade dessa quantidade de informação foi, e continua sendo, um fator de grande importância para o desenvolvimento de estudos na área de genômica dos procariotos na bioinformática.

Sabendo que o DNA tem uma estrutura baseada em duas fitas ligadas, podemos definir o tamanho da molécula pelo número de pares presentes, que será o mesmo número de nucleotídeos presentes em uma única fita. Para as bactérias, o tamanho de genoma geralmente varia em entre 1.000.000 e 10.000.000 pares de bases. O genoma humano, por sua vez, apresenta cerca de 3.000.000.000 de pares (WATSON et al., 2004). Podemos afirmar, então, que o tamanho reduzido dos genomas dos procariotos facilita a aplicação de técnicas de sequenciamento, além de simplificar processos de investigação do DNA, pelo simples fato de haver menos informação a ser processada e por causa da simplicidade na maquinaria genética e de seus processos (ALBERTS et al., 2014).

A transferência de material genético dos seres vivos ocorre normalmente de forma vertical, onde um organismo passa a informação para seus descendentes. Entretanto, especialmente nos procariotos, é muito comum a transferência genômica horizontal (TGH), onde a transferência de informação ocorre independente da reprodução do organismo (LAWRENCE; ROTH, 1996). As bactérias apresentam casos de transferência genômica horizontal com maior frequência em relação a outros organismos mais complexos, sendo este o principal fator para o desenvolvimento de resistência a antibióticos, além de exercer um importante papel na evolução desse seres (GYLES; BOERLIN, 2014).

O material transferido horizontalmente, quando já localizado no genoma do organismo receptor, é o que define a ilha genômica, que nada mais é do que uma parte do genoma que apresenta origens horizontais. As IGs, por sua vez, podem apresentar diferentes tamanhos, variando normalmente entre 10.000 e 200.000 nucleotídeos. Este é um fator que dificulta significativamente a sua identificação, obrigando o método de detecção a se adaptar a suas diferentes possibilidades de tamanho (HACKER; KAPER, 2000).

2.1 TIPOS DE MÉTODOS PARA IDENTIFICAÇÃO DE ILHAS GENÔMICAS

Sobre os métodos utilizados na identificação de IGs, existem dois tipos principais: métodos baseados em comparação e métodos baseados em composição.

Os métodos baseados em comparação se aproveitam do fato de que duas espécies próximas na árvore filogenética apresentam materiais genéticos muito parecidos entre si. Assim, é possível realizar a comparação do genoma a ser investigado com genomas de outras espécies próximas (genomas de referência). Caso sejam identificadas porções do genoma investigado que não estão presentes em nenhum dos genomas das outras espécies, é de grande probabilidade que essas porções sejam IGs (LANGILLE; HSIAO; BRINKMAN, 2010).

Uma das principais limitações dos métodos baseados em comparação é a escolha dos genomas de referência. Caso sejam selecionadas espécies que não sejam filogeneticamente próximas o suficiente do organismo investigado, é possível que IGs sejam identificadas erroneamente. Isto se dá quando o genoma investigado apresenta genes que não são presentes nos genomas escolhidos, não porque estes genes são resultados de TGH, mas sim porque a própria distância filogenética entre os organismos justifica essa ausência (LANGILLE; HSIAO; BRINKMAN, 2010).

Os métodos baseados em composição, por sua vez, não apresentam a necessidade da presença de genomas adicionais para que as IGs possam ser identificadas. Isto acontece porque tais métodos utilizam o próprio genoma investigado como referência para determinar a existência de porções de DNA que fogem do padrão esperado. Estes métodos procuram estabelecer um fator que possa representar o genoma em sua totalidade e, a partir disto, comparar parcelas do genoma a esse fator representativo. Assim, as parcelas que se mostrarem mais divergentes, após processo de comparação, são aquelas que apresentam mais chances de serem resultantes de TGH.

O atributo mais comumente utilizado para a comparação, nos métodos baseados em composição, é a quantidade de GC (guanina e citosina) presente no genoma, chamada normalmente de "conteúdo GC". Nessa abordagem, o genoma é dividido em várias partes, chamadas de "janelas". Então, é calculada a frequência de guanina e citosina em cada uma dessa janelas para que, após o cálculo de todas as frequências, seja possível comparar o valor de cada uma com a média total do genoma (KARLIN, 2001). De maneira similar, é possível realizar a mesma abordagem levando também em conta os valores de adenina e timina presentes no DNA.

2.2 Métodos de Clusterização

A bioinformática tem como principal objetivo a manipulação e interpretação de informações biológicas através da aplicação de métodos da matemática, estatística

e computação (COHEN, 2015). Podemos afirmar, então, que é possível aplicar métodos de mineração de dados e aprendizagem de máquina em dados de origem biológica, como genomas, por exemplo. Na literatura, é possível encontrar a utilização de métodos de clusterização para a identificação de IGs em DNAs de bactérias (DE BRITO, 2016).

Na área de aprendizagem de máquina, existem dois tipos de aprendizagem: supervisionada e não-supervisionada. No primeiro tipo, existe um conjunto de dados de entrada, juntamente com um conjunto de saídas esperadas, por exemplo: dadas as características de um tumor cancerígeno (entrada), identificar se é benigno ou maligno (saída). No segundo tipo, não existem saídas previstas (BOUSQUET et al., 2004). Um bom exemplo para esse caso são problemas de agrupamento (clusterização), como é o caso do que está sendo apresentado neste trabalho: dado um conjunto de dados, divida-o em vários grupos, de forma que um objeto seja sempre mais semelhante aos outros objetos presentes no seu grupo do que aos objetos presentes nos grupos restantes (BAILEY, 1994). DE BRITO (2016) fez uso do algoritmo de clusterização *mean shift* para a identificação de IGs. Para entender o funcionamento do *mean shift,* imaginemos um plano. Inicialmente várias janelas são criadas neste plano, cada uma com um ponto médio. Assim, baseado na proximidade de cada ponto do banco de dados com cada um dos pontos médios, esse ponto é absorvido pela janela, fazendo com que ela aumente e que um novo ponto médio seja calculado. Este processo se repete até que haja convergência, ou seja, não ocorra mais alterações nas posições das janelas.

O *Expectation-Maximization* (EM) é um algoritmo iterativo que tem como objetivo encontrar valores de máxima verossimilhança de parâmetros em modelos probabilísticos que dependem de variáveis não-observáveis. O processo de iteração alterna entre o passo E (*expectation),* onde é realizada uma estimativa da função log-verossimilhança com base nos valores dos parâmetros atuais. O passo M (*maximization)* calcula novos valores para os parâmetros levando em conta a maximização da estimativa da função obtida no passo anterior (DEMPSTER, 1977).

Para fins didáticos, o algoritmo será apresentado através de um exemplo prático, a seguir. Dado o conjunto de dados unidimensionais abaixo (Figura 2), temos como objetivo realizar o agrupamento desses dados em dois grupos diferentes.

Figura 2. Conjunto de dados unidimensionais.

Fonte: https://goo.gl/SbnMB3

Inicialmente, duas funções distribuição de probabilidade normal (Figura 3) são posicionadas aleatoriamente sobre o conjunto de dados.

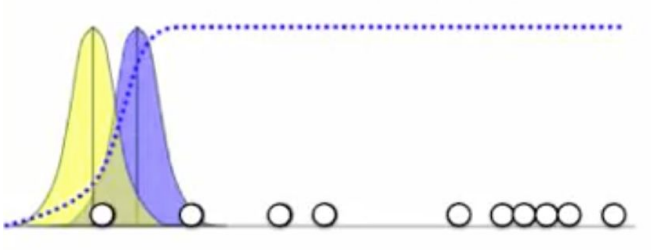

Figura 3. Posição inicial das distribuições normais em relação ao conjunto de dados.

Fonte: https://goo.gl/SbnMB3

Dada a situação inicial, o próximo passo é calcular a probabilidade de pertencimento à distribuição amarela ou azul, para cada um dos pontos. Para isso, utilizamos a seguinte função (Equação 1):

$$P(x_i \mid b) = \frac{1}{\sqrt{2\pi\sigma_b^2}} \exp\left(-\frac{(x_i - \mu_b)^2}{2\sigma_b^2}\right)$$

Equação 1. Cálculo da probabilidade de cada ponto do conjunto de dados.

Fonte: https://goo.gl/SbnMB3

Na função acima, o valor de x representa a posição do ponto no eixo unidimensional. A média da função de distribuição é representada por μ, e a variância

pela letra σ. Após calculados os valores iniciais das probabilidades, temos a seguinte situação (Figura 4):

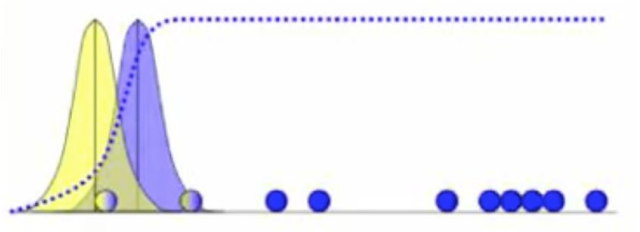

Figura 4. Resultados iniciais dos cálculos de probabilidade.
Fonte: https://goo.gl/SbnMB3

De acordo com a imagem acima, podemos afirmar que o primeiro ponto à esquerda tem maior probabilidade de pertencer à distribuição amarela do que à distribuição azul. Conforme os pontos seguem à direita, a probabilidade de algum deles pertencer à qualquer uma das duas distribuições diminui. Entretanto, como a distribuição azul é aquela que está mais à direita, esses pontos têm maior probabilidade de pertencer à esta distribuição do que à distribuição amarela, e é por isso que estão todos pintados de azul (mesmo havendo ainda uma mínima chance deles pertencerem à distribuição amarela). Agora, temos que atualizar o valor da média e da variância das distribuições de probabilidade. Para a distribuição azul, por exemplo, levamos em conta os pontos que estão assimilados em sua maioria com a cor azul. Abaixo, podemos ver as funções para o cálculo da média e variância (Equações 2 e 3), onde '*b*' representa a probabilidade do ponto ser azul e '*x*' representa a posição do ponto no eixo. Fazemos da mesma forma também para a distribuição amarela.

$$\mu_b = \frac{b_1 x_1 + b_2 x_2 + \ldots + b_n x_{n_b}}{b_1 + b_2 + \ldots + b_n}$$

$$\sigma_b^2 = \frac{b_1 (x_1 - \mu_1)^2 + \ldots + b_n (x_n - \mu_n)^2}{b_1 + b_2 + \ldots + b_n}$$

Equações 2 e 3. Fórmulas usadas na atualização dos valores da média e variância das distribuições de probabilidade, respectivamente.

Fonte: https://goo.gl/SbnMB3

Após o cálculo da nova média e variância, obtemos as distribuições (Figura 5):

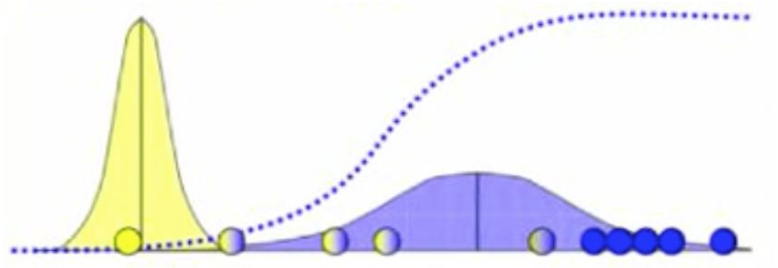

Figura 5. Distribuições de probabilidade após a primeira iteração do EM.

Fonte: https://goo.gl/SbnMB3

O EM irá repetir este processo de estimativa de parâmetros (*expectation*) e cálculo de novas distribuições (*maximization*) até que os resultados venham a convergir. Ao final da execução do método, o resultado estará da seguinte forma (Figura 6):

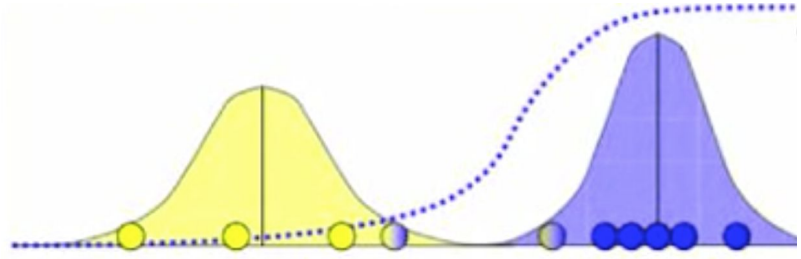

Figura 6. Resultado final da execução do EM para o exemplo.

Fonte: https://goo.gl/SbnMB3

Assim, os pontos que estão mais próximos da distribuição azul são agrupados todos em um *cluster*. Semelhantemente, os pontos mais próximos da distribuição amarela são agrupados em outro *cluster*. Em suma, o resultado final é obtido, tendo em vista que os pontos foram agrupados em dois *clusters* diferentes de acordo apenas com suas posições no eixo uni-dimensional.

3. METODOLOGIA

O trabalho aqui presente propõe a aplicação do EM para a identificação de ilhas genômicas em organismos procariotos. Um dos fatores mais importantes para a escolha deste método é que não é necessária a definição prévia do número de *clusters* a serem gerados pelo método, pois este valor pode ser calculado por validação-cruzada antes da execução do EM. Isto se faz necessário pelo motivo de que, quando se deseja detectar IGs em um certo DNA, não se sabe quantas ilhas existem, ou até mesmo se alguma ilha existe nesse genoma.

Validação-cruzada é uma técnica utilizada para estimar a capacidade de generalização de um modelo, a partir de um conjunto de dados. A idéia central desta técnica consiste na divisão do conjunto de dados em subconjuntos mutuamente exclusivos, onde alguns subconjuntos são utilizados como dados de treinamento do modelo e os outros são utilizados como dados para a validação (teste) do modelo gerado. Assim, é possível determinar se este modelo irá responder de forma esperada aos dados de entrada das mais variadas naturezas (KOHAVI, 1995).

O algoritmo foi implementado utilizando a linguagem de programação Java, tendo como auxílio as bibliotecas disponibilizadas pelo software Weka (HALL et al, 2009), que apresenta um conjunto de algoritmos de mineração de dados já implementados, inclusive o EM. Assim, foi necessário apenas definir o valor dos parâmetros requisitados pelo método.

Ao estudar os melhores valores para os parâmetros, obtivemos os seguintes:

- O valor do número de *clusters* não é dado, pois é definido por validação cruzada antes da execução do método de clusterização.
- O número de *folds* utilizado na validação cruzada é fixado em 10, conforme discutido na literatura (KOHAVI, 1995). Este valor irá

definir o número de partes em que o conjunto de dados analisado pelo EM será dividido. Neste caso (10-*fold),* 9 partes do conjunto serão utilizadas para treinamento do modelo e 1 parte será utilizada para teste. O algoritmo, então, é executado 10 vezes, assegurando que todas as partes foram utilizadas como conjunto de teste. Finalizando esta etapa, normalmente se calcula uma média dos resultados obtidos em cada um dos 10 testes, obtendo o resultado final. Este procedimento permite uma avaliação mais completa do funcionamento do algoritmo (EM, neste caso), pois utiliza todas as combinações possíveis do conjunto de dados, segundo o número de partes em que o conjunto foi dividido, e executa o método para cada uma dessas combinações.

- O melhor valor possível da semente, que é um número aleatório que irá indicar as posições iniciais das distribuições de probabilidade, foi calculado da seguinte forma: o EM foi executado 100 diferentes vezes, cada uma com um valor diferente de semente (gerado por uma função pseudo-aleatória). Após as 100 execuções, a semente que gerou a identificação do maior número de ilhas foi selecionado. A quantidade de 100 execuções com valores de semente diferentes foi escolhida porque, para números maiores, o resultados permaneceram inalterados e, para números menores, os resultados eram menos interessantes do que para esse valor.

- Número máximo de *clusters*: não limitado.

- Número máximo de iterações realizadas pelo EM: não limitado.

- Número de *threads* de execução paralela do EM: não limitado.

- Menor alteração no valor da função log-verossimilhança necessário para se realizar uma nova iteração do EM: 0. Ao definir este valor, devemos ter em mente que o EM deve parar sua execução quando o valor da função convergir, ou seja, não mais se alterar. Por isso a

escolha do valor 0. O método EM, por definição, atinge convergência em algum momento (WU, 1983).

- Menor alteração no valor da função log-verossimilhança da validação cruzada necessário para considerar o aumento no número de *clusters* a serem definidos: pelo mesmo motivo do parâmetro anterior, foi definido em 0.

Para medir a eficácia da solução proposta neste trabalho, o método foi testado em um conjunto de bactérias que tiveram seus genomas já analisados por outros autores (ZHANG, R.; ZHANG, C., 2004; ZHANG, C.; ZHANG, R, 2004; WATERHOUSE; RUSSELL, 2006; CHATTORAJ et al., 2010; OCHMAN; LAWRENCE; GROISMAN, 2000; DE BRITO et al., 2016). Os arquivos .fastq utilizados nos testes foram obtidos através do banco de dados do NCBI de genomas de bactérias já sequenciados (ftp.ncbi.nih.gov/genomes/Bacteria/) e foram adaptados em seguida para arquivos .arff, que é o formato requerido pelos métodos implementados nas bibliotecas do software Weka (HALL et al, 2009). Abaixo podemos visualizar a lista dos arquivos .fastq baixados do banco de dados do NCBI.

Tabela 1. Organismos selecionados para os testes do EM.

Genoma
Corynbacterium Glutamicum ATCC 13032
Vibrio Vulnificus CMCP6 chromosome I
Rhodopseudomonas palustris CGA009
Streptococcus mutans UA159
Vibrio cholerae chromosome II
Vibrio Vulnificus YJ016 chromosome I
Mycoplasma Genitalium G37
Rickettsia prowazekii str. Breinl

Fonte: Elaborada pelo autor.

O genoma de menor tamanho, dentre os listados acima, é o do organismo *Mycoplasma Genitalium,* com 0,58Mb (megabases) de tamanho. Já o maior genoma é o da bactéria *Rhodopseudomonas palustris,* com 5,45Mb de tamanho.

O tamanho da janela utilizada foi de 50kb, como utilizado por outros autores (DE BRITO, et al., 2016). Assim, sabendo também que o tamanho típico das IGs varia entre 10kb e 200kb (HACKER; KAPER, 2000), podemos definir que, para o EM, uma ilha genômica será mostrada como um conjunto de uma à cinco janelas agrupadas em um único *cluster* diferente do *cluster* que abriga a maioria das janelas. Após a definição do tamanho das janelas, foi calculada a frequência de cada um dos nucleotídeos (A, T, C e G) em cada janela. Estes valores definem a posição da janela em um espaço de quatro dimensões (uma dimensão para cada nucleotídeo). Baseado nessas posições, o EM será capaz de calcular a similaridade das janelas entre si.

3.1 O FORMATO FASTQ

Os arquivos de extensão .fastq constituem, atualmente, o principal tipo de arquivo responsável por armazenar genomas, além de outras sequências biológicas. O formato fastq permite também armazenar a pontuação de qualidade referente a cada item da sequência. Esta pontuação serve de indicador da probabilidade de um item da sequência ter sido lido erroneamente durante o sequenciamento (COCK, 2009).

A primeira linha do arquivo é responsável por expor a identificação da informação sequenciada, seguida da segunda linha, que é responsável pela sequência propriamente dita. A terceira linha contém algum possível comentário e a quarta, e última, linha contém as pontuações de qualidade. Para o problema de identificação de IGs, é necessário levar em consideração apenas a linha que contém a sequência de bases nitrogenadas (A, T, C ou G), pois é a partir dela que obtemos a frequência de cada uma das bases no genoma completo, que é o parâmetro usado para identificar as ilhas.

O padrão de quatro linhas se repete até o fim da sequência ser atingido (COCK, 2009). Veja um exemplo a seguir (Figura 7):

@ID_DA_SEQUENCIA
GATTTGGGGTTTAAATCCATTTGTTCAACAAATAGTAAATCCATTTGTTCAAC
+
!''((((***+))%%%++)(%%%%).1***-+*''))**55%%%).1***-+*''))**C65*

Figura 7: Exemplo de arquivo .fastq.

Fonte: Elaborada pelo autor.

Com isso em mente, podemos discutir os resultados obtidos.

4. APRESENTAÇÃO E ANÁLISE DOS RESULTADOS

Os resultados gerados pelo EM serão apresentados a seguir, sempre acompanhados dos resultados obtidos por dois outros métodos da literatura. As ilhas identificadas por estes métodos são os parâmetros utilizados neste trabalho para medir a eficácia do método aqui apresentado.

É preciso ressaltar que os organismos *Rickettsia prowazekii* e *Mycoplasma Genitalium* não apresentaram nenhuma IG identificada pelo EM e nem por outros métodos na literatura. Abaixo, podemos analisar o restante das bactérias analisadas.

Para o organismo *Corynbacterium Glutamicum ATCC 13032*, a IG identificada pelos outros métodos também foi detectada pelo EM. Além disso, outras 4 ilhas também foram detectadas.

Tabela 2. Resultados do EM para *Corynbacterium Glutamicum ATCC 13032.*

Número da IG	*Mean Shift - BRITO, D. M. de (2016)* Posição da janela (Mb)	Zhang, R. e Zhang, C. (2004) Posição da IG (Mb)	Expectation - Maximization Posição da janela (Mb)
1	———	———	0,350 — 0,400
2	———	———	0,600 — 0,650
3	———	———	1,100 — 1,150
4	———	———	1,400 — 1,450
6	1,800 — 2,000	1,776 — 1,987	1,750 — 2,000

Fonte: Elaborada pelo autor.

Para a bactéria *Vibrio vulnificus CMCP6 chromosome I*, duas IGs identificadas estão de acordo com a literatura. Além dessas, mais duas também foram

detectadas. Abaixo, estão listadas as tabelas com os resultados referentes às bactérias restantes.

Tabela 3. Resultados do EM para *Vibrio vulnificus CMCP6 chromosome I.*

Número da IG	*Mean Shift - BRITO, D. M. de (2016)* Posição da janela (Mb)	Zhang, R. e Zhang, C. (2004) Posição da IG (Mb)	Expectation - Maximization Posição da janela (Mb)
1	0,350 — 0,400	0,355 — 0,395	0,350 — 0,400
2	———	———	0,750 — 0,800
3	1,000 — 1,050	———	———
4	———	———	1,600 — 1,650
5	2,450 — 2,650	2,438 — 2,605	2,450 — 2,600
6	———	———	2,650 — 2,700
7	———	3,248 — 3,281	———

Fonte: Elaborada pelo autor.

Tabela 4. Resultados do EM para *Rhodopseudomonas palustris CGA009.*

Número da IG	*Mean Shift - BRITO, D. M. de (2016)* Posição da janela (Mb)	Zhang, R. e Zhang, C. (2004) Posição da IG (Mb)	Expectation - Maximization Posição da janela (Mb)
1	1,150 — 1,200	———	———
2	1,250 — 1,300	———	———
3	2,500 — 2,550	2,481 — 2,564	———
4	3,650 — 3,700	———	———
5	3,750 — 3,800	3,729 — 3,807	3,750 — 3,800
6	3,950 — 4,000	———	———
7	4,550 — 4,700	4,578 — 4,678	4,550 — 4,650
8	5,200 —5,250	———	———

Fonte: Elaborada pelo autor.

Tabela 5. Resultados do EM para *Streptococcus mutans UA159.*

Número da IG	*Mean Shift - BRITO, D. M. de (2016)* Posição da janela (Mb)	*Waterhouse e Russell (2006), Chattoraj et al. (2010)* Posição da IG (Mb)	Expectation - Maximization Posição da janela (Mb)
1	1,250 — 1,300	1,250 — 1,300	1,250 — 1,300

Fonte: Elaborada pelo autor.

Tabela 6. Resultados do EM para *Vibrio cholerae chromosome II.*

Número da IG	*Mean Shift - BRITO, D. M. de (2016)* Posição da janela (Mb)	*Nag et al. (2006)* Posição da IG (Mb)	Expectation - Maximization Posição da janela (Mb)
1	0,300 — 0,450	0,302 — 0,436	0,300 — 0,450
2	———	———	0,800 — 0,850

Fonte: Elaborada pelo autor.

Tabela 7. Resultados do EM para *Vibrio vulnificus YJ016 chromosome I.*

Número da IG	*Mean Shift - BRITO, D. M. de (2016)* Posição da janela (Mb)	*Nag et al. (2006)* Posição da IG (Mb)	Expectation - Maximization Posição da janela (Mb)
1	———	0,159 — 0,167	———
6	1,800 — 1,950	1,757 — 1,936	1,800 — 1,950
7	2,200 — 2,250	———	2,200 — 2,300

Fonte: Elaborada pelo autor.

Podemos observar, a partir dos resultados acima, que a maioria das IGs identificadas simultaneamente pelos outros métodos presentes na literatura também foram identificadas pelo EM. Apenas uma IG que foi identificada pelos outros dois autores utilizados na comparação não foi identificada pelo EM (a IG posicionada

entre 2,500Mb e 2,550Mb do organismo *Rhodopseudomonas palustris*), o que corrobora a eficiência do método apresentado neste trabalho.

A região entre 2,500Mb e 2,550Mb do organismo *Rhodopseudomonas palustris* é constituída principalmente por proteínas conjugativas, essenciais para o processo de conjugação bacteriana, processo sexual de transferência de genes de uma bactéria doadora para uma receptora. Dessa forma, essa sequência de DNA traz características de ilha genômica horizontal, sendo necessário o estudo biológico mais detalhado sobre características da região, como outros elementos presentes e idade desses genes (DE BRITO et al., 2016).

Além disso, é importante destacar a descoberta de 4 ilhas genômicas não descritas por nenhum dos outros dois autores aqui referenciados em *Corynbacterium Glutamicum ATCC 13032*, além de 3 ilhas em *Vibrio vulnificus CMCP6 chromosome I* e 1 em *Vibrio cholerae chromosome II*. Novamente, a mesma análise biológica se faz necessária nessas regiões gênicas, buscando características que possam justificar sua detecção pelo método. Lembrando ainda que alguns elementos gênicos, como proteínas ribossômicas, induzem a caracterização de ilhas genômicas por possuírem um viés para o uso de códons muito maior que os outros genes nos genomas (ZHANG E ZHANG, 2005).

O estudo, portanto, mostra a importância de usar diferentes métodos de bioinformática na busca por ilhas genômicas, como também de variações do mesmo método, como os baseados na assinatura gênica, podendo sugerir várias regiões em comum, como ilhas (e outras tantas diversas) que podem tratar de regiões resultantes de transferência gênica horizontal.

5. CONCLUSÕES E TRABALHOS FUTUROS

A eficácia do método proposto foi confirmada através de testes utilizando um conjunto de genomas de oito organismos diferentes. A comparação dos resultados obtidos pelo EM com os resultados obtidos por outros autores foi de suma importância para a corroboração do método apresentado neste trabalho. Entretanto, a aplicação do EM apresenta alguma limitações que motivam trabalhos futuros.

A existência de ilhas identificadas pelo EM que não foram documentadas previamente faz que seja necessária uma justificação, possivelmente de cunho biológico, para a identificação desses trechos de genoma como IGs.

Outro possível trabalho futuro a ser realizado é a alteração do tamanho da janela utilizada para calcular a frequência de cada base nitrogenada presente no genoma (seção 2.1). Esta alteração poderá gerar novos valores de frequência para cada janela e, consequentemente, novos resultados no processo de clusterização. Além disso, também é possível utilizar a frequência das bases adenina e guanina no cálculo da frequência, mesmo sabendo o "conteúdo GC" é mais comumente utilizado, como explicado na seção 2.1. Caso sejam utilizadas todas as quatro bases, é possível calcular a ocorrência de várias combinações possíveis entre elas ("ATCGA", "CAGT", "ATG", entre outras) e comparar a frequência de cada uma dessas combinações em cada janela (ou seja, da mesma que se utilizou o conteúdo GC).

Tendo em vista os resultados obtidos, podemos concluir que o EM constitue um método eficiente para a identificação IGs, quando comparado com os outros métodos da literatura. Além disso, após a implementação dos trabalhos futuros aqui indicados, é possível que o método seja capaz de obter, num futuro próximo, resultados ainda mais próximos do ideal.

6. REFERÊNCIAS

ALBERTS, Bruce et al. Molecular Biology of the Cell (6th ed.), 2014.

BAILEY, Ken. "Numerical Taxonomy and Cluster Analysis". Typologies and Taxonomies. p. 34, 1994.

BOUSQUET, Olivier et al. (2004). Advanced Lectures on Machine Learning. 2004.

CHATTORAJ, Partho et al. ClpP of Streptococcus mutans differentially regulates expression of genomic islands, mutacin production, and antibiotic tolerance. Journal of bacteriology, v. 192, n. 5, p. 1312-1323, 2010.

COCK, Peter et al. "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants". Nucleic Acids Research. 38 (6): 1767–1771, 2009.

COHEN, Jacques. Bioinformatics: an Introduction for Computer Scientists. ACM Comput, 2015.

DE BRITO, DANIEL M. et al . A Novel Method to Predict Genomic Islands Based on Mean Shift Clustering Algorithm. Plos One , v. 11, p. e0146352, 2016.

DEMPSTER, Arthur et al. "Maximum Likelihood from Incomplete Data via the EM Algorithm". Journal of the Royal Statistical Society, Series B. 39 (1): 1–38, 1977.

FLEISCHMANN, Robert D. et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science, v. 269, n. 5223, p. 496-512, 1995.

GYLES, C.; BOERLIN, P. "Horizontally transferred genetic elements and their role in pathogenesis of bacterial disease". Veterinary Pathology, p. 51, 2014.

HACKER, Jörg; KAPER, James B. Pathogenicity islands and the evolution of microbes. Annual Reviews in Microbiology, v. 54, n. 1, p. 641-679, 2000.

HALL, Mark et al. The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter, v. 11, n. 1, p. 10-18, 2009.

JUHAS, Mario et al. Genomic islands: tools of bacterial horizontal gene transfer and evolution. FEMS microbiology reviews, v. 33, n. 2, p. 376-393, 2009.

KARLIN, Samuel. Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. Trends in microbiology, v. 9, n. 7, p. 335-343, 2001.

KERFELD, Cheryl et al. "Protein structures forming the shell of primitive bacterial organelles". Science, 2005.

KOHAVI, Ron. "A study of cross-validation and bootstrap for accuracy estimation and model selection". Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, 1995.

LANGILLE, Morgan GI; HSIAO, William WL; BRINKMAN, Fiona SL. Detecting genomic islands using bioinformatics approaches. Nature Reviews Microbiology, v. 8, n. 5, p. 373-382, 2010.

LAWRENCE, Jeffrey G.; ROTH, John R. Selfish operons: horizontal transfer may drive the evolution of gene clusters. Genetics, v. 143, n. 4, p. 1843-1860, 1996.

MADIGAN, Michael T. et al. Brock Biology of microorganisms. Pearson, 2015.

OCHMAN, Howard; LAWRENCE, Jeffrey G.; GROISMAN, Eduardo A. Lateral gene transfer and the nature of bacterial innovation. Nature, v. 405, n. 6784, p. 299-304, 2000.

SANGER, Frederick et al. "Nucleotide sequence of bacteriophage phi X174 DNA". Nature. 265 (5596): 687–95, 1977.

SILVA, César et al. Biologia: Volume Único. 6ª Edição. Saraiva, 2015.

WANG, Yuja; BRYANT, Samuel. The NCBI handbook, 2nd edition, 2014.

WATERHOUSE, Janet C.; RUSSELL, Roy RB. Dispensable genes and foreign DNA in Streptococcus mutans. Microbiology, v. 152, n. 6, p. 1777-1788, 2006.

WATSON, James et al. “Ch9-10”, Molecular Biology of the Gene, 5th ed., Pearson Benjamin Cummings; CSHL Press, 2004.

WU, Jeff. "On the Convergence Properties of the EM Algorithm". Annals of Statistics, 1983.

ZHANG, Chun-Ting; ZHANG, Ren. Genomic islands in Rhodopseudomonas palustris. Nature biotechnology, v. 22, n. 9, p. 1078-1079, 2004.

ZHANG, Ren; ZHANG, Chun-Ting. A systematic method to identify genomic islands and its applications in analyzing the genomes of Corynebacterium glutamicum and Vibrio vulnificus CMCP6 chromosome I. Bioinformatics, v. 20, n. 5, p. 612-622, 2004.