UNIVERSIDADE FEDERAL DE PELOTAS

Programa de Pós-Graduação em Biotecnologia

Dissertação

Aplicação de algoritmos de mineração de dados para classificação molecular de *Leptospira* spp.

Julia Labonde

Pelotas, 2016

Julia Labonde

Aplicação de algoritmos de mineração de dados para classificação molecular de *Leptospira* spp.

Dissertação apresentada ao Programa de Pós-Graduação em Biotecnologia da Universidade Federal de Pelotas, como requisito parcial à obtenção do título de Mestre em Ciências. Área do Conhecimento: Biologia Molecular e Bioinformática.

Orientador: Odir Antônio Dellagostin

Pelotas, 2016

Universidade Federal de Pelotas / Sistema de Bibliotecas
Catalogação na Publicação

L122a Labonde, Julia

Aplicação de algoritmos de mineração de dados para classificação molecular de leptospira spp. / Julia Labonde ; Odir Antônio Dellagostin, orientador. — Pelotas, 2016.

37 f. : il.

Dissertação (Mestrado) — Programa de Pós-Graduação em Biotecnologia, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, 2016.

1. Leptospira. I. Dellagostin, Odir Antônio, orient. II. Título.

CDD : 570.285

Elaborada por Maria Beatriz Vaghetti Vieira CRB: 10/1032

BANCA EXAMINADORA

Prof. Dr. Luciano Carlos da Maia (FAEM, UFPel)

Prof. Dr. Luciano da Silva Pinto (CDTec, UFPel)

Prof. Dr[a]. Alan McBride (CDTec, UFPel)

Prof. Dr. Odir Antônio Dellagostin (Orientador – CDTec, UFPel)

Agradecimentos

Primeiramente agradeço à minha mãe “Tuti” por todo apoio emocional, psicológico e motivacional durante estes dois anos de aprendizado. Dentro do espectro familiar prolongo os agradecimentos à minha irmã e melhor amiga Maria Luyza, que, apesar de fisicamente parecer-se muito comigo, tem personalidade bastante distinta, o que nos proporciona experiências engrandecedoras.

Muito obrigada aos colegas do laboratório de Bioinformática Marcus Eslabão, Frederico Kremer e Paulo Ricardo Porto por todo o aprendizado, piadas, zoeira e conversas sem fundamento que deixaram o ambiente de trabalho leve e gostoso de trabalhar. Obrigado aos demais estagiários e pós-graduandos, que, apesar de pertencerem a outros projetos, vivenciaram momentos de aprendizado e amizade junto comigo.

Agradeço ao meu orientador, Odir Dellagostin, por todos os ensinamentos durante todo este período.

Agradeço também a Capes por ter disponibilizado minha bolsa de mestrado e demais recursos para a manutenção e execução deste projeto.

Além de todas essas pessoas a quem devo agradecimentos especiais, agradeço fortemente ao meu marido e melhor amigo Juliano Stefan por aturar meu comportamento birrento e as loucuras da minha cabeça, mas principalmente pela paciência e amor dedicados a mim nesses quase sete anos juntos.

Resumo

LABONDE, Julia. **Aplicação de algoritmos de mineração de dados para classificação molecular de *Leptospira* spp.** 2016. 37f. Dissertação (Mestrado) - Programa de Pós-Graduação em Biotecnologia. Universidade Federal de Pelotas, Pelotas.

A leptospirose é uma doença infecciosa de importância mundial, que afeta humanos e animais, causada por espiroquetas patogênicas pertencentes ao gênero *Leptospira*. Para a área epidemiológica e clínica é fundamental que os laboratórios tenham a capacidade de identificar e classificar com precisão as espécies de *Leptospira* que causam doença, para que sejam tomadas decisões coerentes com relação à saúde pública. Neste estudo, nós relatamos pela primeira vez a utilização de ferramentas de mineração de dados para fins de classificação de cepas do gênero *Leptospira*. Vinte e cinco *loci* referentes a 15 genes foram selecionadas e analisados em 600 genomas rascunho de *Leptospira,* com o propósito de buscar polimorfismos que pudessem ser utilizados na classificação de cada espécie. Para isso, foram utilizados os algoritmos baseados em mineração de dados C4.5, Naive Bayes e *Support Vector Machine*. Todos os algoritmos computacionais de mineração de dados utilizados neste trabalho apresentaram valores de acurácia acima de 93% para classificação de *Leptospira* a nível de espécie, no entanto, o algoritmo C4.5, além de atingir a melhor acurácia de classificação (95.6%), também apresentou os genes que contribuíram para o resultado final da análise. O mesmo banco de dados genômicos utilizado pelos algoritmos computacionais foi submetido a testes com a metodologia MLST – técnica mais utilizada para classificação molecular de espécies deste gênero – no entanto, nenhum dos testes apresentou acurácia superior a 80%. Visto o algoritmo de mineração de dados C4.5 atingir uma acurácia superior aos outros algoritmos, pode-se concluir que C4.5 é uma ferramenta de mineração de dados bastante promissora para classificar espécies de *Leptospira*.

Palavras-chave: *Leptospira*, Classificação, Mineração de Dados, algoritmos computacionais.

Abstract

LABONDE, Julia. Application of data mining algorithms for molecular classification of *Leptospira* spp. 2016. 37f. Thesis (Master's degree) - Graduate Program in Biotechnology. Federal University of Pelotas, Pelotas.

Leptospirosis is an infectious disease of global importance that affects humans and animals caused by pathogenic spirochetes belonging to the genus *Leptospira*. For epidemiological and clinical areas, it is essential that laboratories have the ability to identify and classify accurately species of *Leptospira* that cause disease, to take decisions consistent with respect to public health. In this study, we report for the first time the use of data mining tools for the purposes of strain classification of the genus *Leptospira*. Twenty-five loci related to 15 genes were selected and analyzed in 600 *Leptospira* draft genomes in order to search polymorphisms that could be used for the classification of each species. For this, data mining-based algorithms - C4.5, Naive Bayes and SVM - were used. All data mining computational algorithms used in this study showed accuracy levels above 93% for *Leptospira* classification species, however, the C4.5 algorithm achieve the best accuracy rating (95.6%) and presented the genes that contributed to the final result of the analysis. The same genomic database used by computer algorithms has been tested with the MLST methodology – most used technique for molecular classification of species of this genus - however, none of the tests show accuracy higher to 80%. Because data mining algorithm C4.5 achieve better accuracy than other algorithms, it can be concluded that C4.5 is a very promising data mining tool to classify species of *Leptospira*.

Keywords: *Leptospira*, Classification, Data Mining, Computational Algorithms.

Lista de Figuras

Figura 1. Árvore de decisão gerada pelo algoritmo C4.5 para classificar espécies de Leptospira.....29

Figura 2. Classificação de Leptospira spp. predita por algoritmos computacionais baseados em mineração de dados e pela metodologia do MLST.....30

Lista de Tabelas

Tabela 1. *Loci* referentes aos genes dos três esquemas de MLST e genes completos extraídos de Leptospira interrogans sorovar Copenhageni L1-130. 20

Tabela 2. Informações sobre tamanho (em pares de base), posição no genoma e LOCUS_TAG de cada gene utilizado neste trabalho. 21

Tabela 3. Distribuição das sequências genômicas de espécies do gênero Leptospira recuperadas do GenBank.....22

Tabela 4. Exemplo da disposição dos dados genômicos no arquivo CSV. 23

Tabela 5. Matriz de confusão gerada pelo algoritmo C4.5.....26

Tabela 6. Matriz de confusão gerada pelo algoritmo Naive Bayes. 27

Tabela 7. Matriz de confusão gerada pelo algoritmo SVM. 28

Lista de Abreviaturas

MAT – *Microscopic Agglutination Test*

AFLP – A*mplified fragment length polymorphism*

MLVA – *Multilocus Variable Number of Tandem Repeat Analysis*

PFGE – *Pulsed-Field Gel Electrophoresis*

MLST – *Multilocus Sequence Typing*

NCBI – *Nacional Center for Biotechnology Information*

SVM – *Support Vector Machine*

BLAST – *Basic Local Alignment Search Tool*

MUSCLE – *Multiple Sequence Comparison by Log-Expectation*

WEKA – *Waikato Environment for Knowledge Analysis*

ST – *Sequence type*

Sumário

1 INTRODUÇÃO GERAL 11

2 REVISÃO BIBLIOGRÁFICA 13

2.1 A Leptospirose 13

2.2 Classificação sorológica e molecular de *Leptospira* 13

2.3 Sequenciamento de DNA e o desenvolvimento do MLST 14

2.4 Utilização de algoritmos de mineração de dados para classificar *Leptospira* 16

3 HIPÓTESE E OBJETIVOS 19

3.1 Hipótese 19

3.2 Objetivo Geral 19

3.3 Objetivos Específicos 19

4 METODOLOGIA 20

4.1 Criação do Banco de Dados 20

4.2 Aplicação dos algoritmos de mineração 23

4.3 Avaliação das metodologias de classificação 23

5 RESULTADOS 25

5.1 Aplicação dos algoritmos de mineração de dados 25

5.2 Avaliação das metodologias de classificação 30

6 DISCUSSÃO GERAL E PERSPECTIVAS 31

6.1 Comparação dos resultados gerados pelos algoritmos de mineração de dados 31

6.2 Relação entre a metodologia de mineração de dados e a metodologia MLST 32

6.3 Relação entre a mineração de dados e o futuro da classificação de espécies do gênero *Leptospira* 33

7 CONCLUSÃO GERAL 34

8 REFERÊNCIAS 35

1 INTRODUÇÃO GERAL

A leptospirose é uma doença infecciosa de importância mundial que afeta humanos e animais e é causada por espiroquetas patogênicas pertencentes ao gênero *Leptospira* (Levett, 2015). A doença tem ampla distribuição geográfica, ocorrendo em áreas de clima temperado e tropical e em ambientes urbanos e rurais e é transmitida através do contato direto ou indireto com a urina de animais infectados com leptospiras patogênicas (Bharti et al., 2003).

Para a área epidemiológica e clínica é fundamental que os laboratórios tenham a capacidade de classificar e identificar com precisão as cepas de *Leptospira* spp. que causam doença para que sejam tomadas decisões coerentes com relação à saúde pública (Ahmed et al., 2006). Todas as espécies do gênero *Leptospira* foram identificadas e descritas através de abordagens moleculares, por outro lado, a epidemiologia da leptospirose foi construída a partir de abordagens sorológicas, que têm sido tradicionalmente utilizadas para identificar e classificar os isolados de *Leptospira* spp. em sorogrupos e sorovares.

No entanto, traçar uma correlação entre isolados classificados pela abordagem molecular e sorológica não é uma tarefa fácil, pois, a identidade inferida pela abordagem molecular varia em relação a identidade inferida pela abordagem sorológica (Varni et al., 2014). Sendo assim, métodos moleculares de identificação de espécies, como, por exemplo, o sequenciamento de DNA, têm sido utilizados como uma alternativa aos métodos sorológicos (Levett et al., 2006). Como as plataformas de sequenciamento têm se tornado cada vez mais baratas e acessíveis (Buermans & den Dunnen, 2014), o número de genomas de *Leptospira* spp. sequenciado aumentou substancialmente, e mais de 600 genomas rascunho (*draft*s) desta espécie estão disponíveis nos bancos de dados públicos. Toda essa informação disponível em formato digital pode ser utilizada em análises de polimorfismos gênicos para identificação de cepas de *Leptospira* spp. Para isso, é necessário utilizar algoritmos computacionais de mineração de dados. A mineração de dados pode ser eficientemente aplicada em áreas que utilizam grandes quantidades de dados, onde seria difícil a aplicação de abordagens tradicionais como, por exemplo, a comparação manual de sequências genômicas (Hastie et al., 2009). Ela inclui um conjunto de técnicas para processamento de dados que utiliza conhecimentos de diversas áreas como inteligência artificial e estatística e fornece

uma maneira de transformar, de um modo inteligente e automatizado, dados processados em informações compreensíveis (Zhuang et al., 2009). Desta forma, a aplicação de algoritmos computacionais baseados em mineração de dados pode ser utilizada para buscar polimorfismos gênicos que auxiliem na identificação e classificação de cepas de *Leptospira*.

2 REVISÃO BIBLIOGRÁFICA

2.1 A Leptospirose

A leptospirose é uma doença bacteriana que afeta humanos e animais. A infecção ocorre através do contato direto com a urina de animais infectados ou através do contato com ambiente contaminado com as espiroquetas. A bactéria entra no organismo através de cortes ou abrasões na pele ou através das membranas mucosas (boca, nariz e olhos) (Bharti et al., 2003; Levett, 2015).

As espiroquetas estão classificadas na Ordem Spirochaetales, a qual compreende três famílias: Brachyspiraceae, Spirochaetaceae e Leptospiraceae. Leptospiraceae por sua vez possui três gêneros, *Leptonema*, *Turneriella* e *Leptospira,* sendo este último o gênero que abriga as espiroquetas mais estudadas devido a sua importância clínica (Levett, 2015). Atualmente estão descritas 22 espécies para o gênero *Leptospira*, das quais 10 são patogênicas – *L. alexanderi*, *L. alstonii*, *L. borgpetersenii*, *L. interrogans*, *L. kirschneri*, *L. kmetyi*, *L. noguchii*, *L. mayottensis*, *L. santarosai*, *L. weilii* e 5 intermediárias – *L. broomii*, *L. fainei*, *L. inadai*, *L. licerasiae*, *L. wolffii* e 7 saprófitas – *L. biflexa*, *L. idonii*, *L. meyeri*, *L. terpstrae*, *L. vanthielii*, *L. wolbachii*, *L. yanagawae* e (Levett, 2015). O grupo de leptospiras intermediárias causa sintomas clínicos moderados (Kmety & Dikken, 1993).

2.2 Classificação sorológica e molecular de *Leptospira*

Todas as espécies do gênero *Leptospira* foram identificadas e descritas através de abordagens moleculares (Adler, 2014). Por outro lado, a epidemiologia da leptospirose foi construída a partir de abordagens sorológicas, que têm sido tradicionalmente utilizadas para identificar e classificar os isolados de *Leptospira* (Adler, 2014). Estas são baseadas na resposta imunológica do hospedeiro frente a epítopos bacterianos. Com base na sorologia, já foram classificados mais de 300 sorovares, sendo estes agrupados em 25 sorogrupos (Adler, 2014). O método sorológico mais comumente utilizado é o teste de aglutinação microscópica (MAT). No entanto, este teste é demorado, requer equipamentos e laboratório específicos, e exige a manutenção de uma coleção de sorovares de leptospiras, tornando esta

metodologia trabalhosa e restrita a laboratórios de referência (Romero & Yasuda, 2006).

Para estudos clínicos e epidemiológicos, é necessário que os laboratórios consigam identificar e classificar de forma acurada as cepas patogênicas de *Leptospira* para que se possam adotar medidas preventivas e de controle da doença (Ahmed et al., 2006). No entanto, traçar uma correlação entre isolados classificados pela abordagem molecular e pela abordagem sorológica não é uma tarefa fácil, pois, o 'táxon' inferido pela abordagem molecular varia em relação à 'sorovar' inferida pela abordagem sorológica (Varni et al., 2014).

Devido a testes sorológicos serem baseados na resposta imunológica do hospedeiro e lançarem um cenário artificial e não confiável da diversidade de sorovares (Nalam et al., 2010), métodos moleculares de classificação de espécies de *Leptospira* e seus sorovares têm sido utilizados como uma alternativa aos métodos sorológicos (Levett et al., 2006). Estes incluem: *amplified fragment length polymorphism* (AFLP) (Vijayachari et al., 2004), *multilocus variable number of tandem repeat analysis* (MLVA) (Majed et al., 2005), *pulsed-field gel electrophoresis* (PFGE) (Galloway & Levett, 2010), *multilocus sequence typing* (MLST) (Ahmed et al., 2006; Boonsilp et al., 2013; Varni et al., 2014), entre outros. Apesar de serem utilizadas como alternativa aos métodos sorológicos, alguns métodos moleculares ainda sofrem com algumas desvantagens. As técnicas de AFLP e PFGE possuem limitado poder de discriminação e baixa reprodutibilidade e a transferência de dados entre laboratórios é difícil. A desvantagem do método de MLVA, até então, é que só classifica *L. interrogans* (Varni et al., 2014).

2.3 Sequenciamento de DNA e o desenvolvimento do MLST

A associação das limitações na sorologia e das desvantagens das técnicas AFLP, PFGE e MLVA, com a popularização do sequenciamento de DNA, permitiu a ascensão da técnica de classificação molecular MLST, que é baseada na análise de polimorfismos de sequências gênicas (Maiden et al., 1998). O primeiro esquema de MLST foi desenvolvido para *Neisseria meningitidis* (Maiden et al., 1998) e logo foi aceito pela comunidade científica. Posteriormente, o MLST foi aplicado com sucesso para mais de 30 gêneros bacterianos (Belén et al., 2009). Rapidamente o MLST tornou-se o método preferido para classificação de microrganismos devido a seus

múltiplos benefícios. O método proporciona um sistema de tipagem preciso e discriminativo para bactérias e outros organismos, permitindo que os dados de sequência gerados sejam facilmente acessados através da internet (Maiden, 2006).

Em 2006 foi descrito o primeiro esquema de MLST para *Leptospira* baseado na sequência de 6 genes: 4 genes *housekeeping* e 2 genes que codificam proteínas de membrana externa (Ahmed et al., 2006). Devido à significativa divergência genética deste gênero, o que dificulta o desenvolvimento de um esquema único, mais dois esquemas de MLST foram descritos mais tarde (Varni et al., 2014; Thaipadungpanit et al., 2007; Boonsilp et al., 2013). Todos os três esquemas estão disponíveis *online* e podem ser acessados através do link http://pubmlst.org/leptospira/.

O esquema 1 é composto pelas sequências dos seguintes genes: *glmU*, *pntA*, *sucA*, *tpiA*, *pfkB*, *mreA*, *caiB*. Este esquema foi desenvolvido a partir do estudo do genoma de 327 isolados de *Leptospira* e tem a capacidade de classificar até 7 espécies: *L. interrogans*, *L. borgpetersenii*, *L. alexanderi*, *L. kirschneri*, *L. noguchii*, *L. santarosai* e *L. weilii* (Thaipadungpanit et al., 2007; Boonsilp et al., 2013). O esquema 2 é composto pelas sequências dos genes: *adk*, *glmU*, *icdA*, *lipL32*, *lipL41*, *mreA* e *pntA*. Este esquema foi desenvolvido a partir do estudo do genoma de 65 isolados de *Leptospira* e é usado para classificar apenas *L. interrogans* e *L. kirschneri* (Varni et al., 2014). O esquema 3 é composto pelas sequências dos genes: *adk*, *icdA*, *lipL32*, *lipL41*, *rrs2* e *secY*. Este esquema foi desenvolvido a partir do estudo do genoma de 120 isolados de *Leptospira* e também permite a tipagem de *L. interrogans* e *L. kirschneri* (Ahmed et al., 2006).

O principal critério do MLST para seleção dos genes considerados bons classificadores é a sua capacidade de classificar espécies até a nível de sorogrupo. Geralmente os genes selecionados para classificação possuem evolução lenta e são conservados e estáveis dentro de uma espécie, levando a uma alta capacidade de discriminação entre elas (Ahmed et al., 2006). Além dos genes do MLST de *Leptospira*, *ligB*, *rrs*, *secY*, *lipL41* e *rpoB* também possuem alto poder discriminativo e podem ser usados na classificação de *Leptospira* (Postic et al., 2000; Cerqueira et al., 2010).

Atualmente, as plataformas de sequenciamento têm se tornado mais baratas e acessíveis, permitindo sequenciar genomas inteiros em uma velocidade consideravelmente rápida (Buermans & den Dunnen, 2014). O desenvolvimento em

alta resolução da determinação da sequência de nucleotídeos e sua análise permitiu a geração de dados genéticos a partir de qualquer *lócus* do cromossomo de qualquer isolado. Uma das maiores vantagens de se obter dados de sequência de nucleotídeos é que eles podem ser disseminados via internet e podem ser acessados de qualquer lugar (Jolley et al., 2004).

Com essa redução no custo do sequenciamento, o número de genomas rascunho de *Leptospira* aumentou bastante, e mais de 600 genomas podem ser encontrados nos banco de dados públicos, como, por exemplo, DDBJ (http://www.ddbj.nig.ac.jp/), EMBL (http://www.ebi.ac.uk/embl/) e GenBank (http://www.ncbi.nlm.nih.gov/genbank).

Visto a pequena quantidade de genomas disponíveis de *Leptospira* utilizada para desenvolver os três esquemas de MLST frente à grande quantidade de sequências genômicas atualmente disponíveis nos bancos de dados, é fundamental que seja realizada uma nova avaliação que inclua todos os genomas atualmente disponíveis, com o propósito de buscar novos polimorfismos gênicos que possam ser usados para aprimorar as metodologias de classificação, aumentando o poder de classificação das cepas de *Leptospira*.

2.4 Utilização de algoritmos de mineração de dados para classificar *Leptospira*

Para realizar a avaliação de todos os genomas disponíveis nos bancos de dados, é necessário utilizar ferramentas de bioinformática que consigam processar grandes bases de dados biológicos. Até o momento, mais de 600 sequências *draft* de *Leptospira* estão disponibilizadas nos bancos de dados, e cada uma delas contém aproximadamente 4,5 milhões de pares de base (pb), ou seja, cerca de 2,7 bilhões de pares de base estão disponíveis em formato digital, e podem ser processados e analisados para fins de estudo dos polimorfismos gênicos de *Leptospira*.

A manipulação de grandes bases de dados exige métodos computacionais robustos que possam suportar o processamento das informações (Praneenararat et al., 2012). Neste contexto, abordagens computacionais de mineração de dados têm sido desenvolvidas para aprimorar a extração de informações úteis dos grandes volumes de dados (Banerjee et al., 2013), e são eficientemente aplicadas em áreas

que utilizam bases de dados amplas, onde seria difícil a aplicação de abordagens tradicionais (Hastie et al., 2009).

A mineração de dados inclui um conjunto de técnicas de processamento de informações que utiliza conhecimentos de diversas áreas, como inteligência artificial e estatística, e fornece a metodologia e tecnologia necessárias para transformar grandes volumes de dados em informações e conhecimentos úteis (Shah & Kusiak, 2004). A mineração de dados fornece uma maneira de transformar, de um modo inteligente e automatizado, dados processados em informações compreensíveis, e tem sido utilizada como ferramenta de auxílio para o entendimento e resolução de problemas e sua correta tomada de decisão (Zhuang et al., 2009).

As principais atribuições da mineração de dados são: associação, agrupamento, classificação, detecção de anomalias e previsão. A atribuição é escolhida de acordo com os dados a serem analisados e a natureza do problema a tratar (Chen & Liu, 2009). Para este trabalho foi realizado uma mineração de dados com base na classificação de *Leptospira*.

Um modelo de classificação é construído para prever uma classe ou categoria, por exemplo: um oficial de empréstimos bancários necessita analizar seus dados para saber quais os requerentes de empréstimo são "seguros" e quais são “de risco" para o banco. Neste caso, a análise de dados é uma tarefa de classificação, em que um modelo de classificação é construído de modo a prever uma classe (ou categoria) como "seguro" ou "de risco" para os dados de pedido de empréstimo (Han et al., 2012). Neste trabalho, os nucleotídeos de cada gene selecionado para classificação de *Leptospira* formam o banco de dados (atributos) a ser analizado, de forma que as espécies constituem as classes as quais queremos prever.

A classificação de dados é um processo constituído de duas etapas: aprendizado e teste. O aprendizado consiste na construção de um modelo de classificação, conhecido como classificador, onde é realizada uma análise de um conjunto de dados pré-determinado (conjunto de treinamento). As regras de classificação construídas na etapa de aprendizado podem ser utilizadas para classificar amostras de dados de problemas futuros. Na etapa teste, o classificador construído anteriormente é utilizado para classificação propriamente dita, onde prevê a classe/categoria de cada atributo presente no banco de dados. A precisão do teste é dada pela porcentagem de dados que foi corretamente classificado por ele. Caso a

acurácia seja alta, o classificador é considerado eficiente e pode ser utilizado para classificar novos casos. (Han et al., 2012).

Diversas técnicas podem ser utilizadas para construir os classificadores, tais como: árvores de decisão, métodos Bayesianos, separação ótima entre classes, entre outros. Normalmente, algoritmos computacionais baseados em mineração de dados são necessários para realizar o processamento de grandes bases de dados. Os algoritmos C4.5, baseado em árvore de decisão (Salzberg, 1994), Naive Bayes, baseado em métodos Bayesianos (Russell & Norvig, 1995) e Support Vector Machine (SVM), baseado em separação ótima entre classes (Cortes & Vapnik, 1995) são exemplos de algoritmos computacionais bastante utilizados para resolver problemas de classificação.

O algoritmo de classificação C4.5 é baseado na indução da árvore de decisão. Uma árvore de decisão é representada por um fluxograma em forma de árvore onde cada nó interno denota um teste em um atributo, cada ramo denota um resultado do teste e cada folha (ou nó folha) denota uma classe. A árvore de decisão é usada para classificação em classes ou categorias. Em um conjunto de dados onde as classes ainda são desconhecidas, os valores de cada atributo da amostra são testados contra a árvore de decisão e um caminho é traçado a partir da raiz até a folha, que detém a previsão de classe para aquele atributo (Han et al., 2012).

O Naive Bayes, baseado no teorema de Bayes, usa relação de probabilidades. Ele calcula a probabilidade de um atributo desconhecido pertencer a cada uma das classes possíveis, ou seja, tenta predizer a classe mais provável. Ou seja, o algoritmo recebe como entrada um atributo desconhecido (não classificado), e retorna como saída a classe mais provável para este atributo de acordo com cálculos probabilísticos.

O *Support Vector Machine* é um algoritmo de classificação linear binário, e não probabilístico. Ele toma como entrada um conjunto de atributos e prediz, para cada um deles, qual de duas possíveis classes o atributo faz parte. Ele tem como objetivo determinar os limites de decisão que possam produzir uma separação ótima entre classes por meio da minimização dos erros. Essa classificação é baseada no princípio de separação ótima entre classes, tal que se as classes são separáveis, a solução é escolhida de forma a separar ao máximo as classes (Cortes & Vapnik, 1995).

3 HIPÓTESE E OBJETIVOS

3.1 Hipótese

Algoritmos de mineração de dados podem ser utilizados para classificar com precisão espécies de *Leptospira*.

3.2 Objetivo Geral

Avaliar a acurácia de classificação de espécies do gênero *Leptospira* através da aplicação de algoritmos computacionais baseados em mineração de dados.

3.3 Objetivos Específicos

- Desenvolvimento de um banco de dados local com informações de sequências genômicas e também sequências gênicas já utilizadas na classificação molecular de *Leptospira* spp.;
- Buscar polimorfismos nas sequências gênicas selecionadas para a posterior classificação de *Leptospira* spp.;
- Identificar o melhor algoritmo computacional de classificação através de análise de acurácia.

4 METODOLOGIA

4.1 Criação do Banco de Dados

O conjunto de dados utilizados neste trabalho foi constituído por 600 genomas rascunho de *Leptospira* e 25 *loci*, pertencentes a 15 genes. A escolha dos *loci* foi baseada nos *loci* utilizados na metodologia de MLST já estabelecidos para classificação. Além dos *loci* utilizados pela metodologia MLST, sequências gênicas inteiras como *rrs*, *secY*, *rpoB*, *ligB* e *lipL41* também foram utilizadas para a análise de classificação. Os *loci* referentes aos genes dos três esquemas de MLST foram recuperados manualmente de seus respectivos repositórios e as sequências (completas) dos genes *rrs*, *secY*, *rpoB*, *ligB* e *lipL41* foram extraídas manualmente do genoma de *L. interrogans* sorogrupo Icterohaemorrhagiae sorovar Copenhageni cepa Fiocruz L1-130 (Tabela 1). A Tabela 2 mostra o tamanho (em pares de base), a posição e a LOCUS_TAG de cada gene utilizado neste trabalho, com base no genoma de *L. interrogans* sorovar Copenhageni Fiocruz L1-130 (NC_005823.1).

Tabela 1. *Loci* referentes aos genes dos três esquemas de MLST e genes completos extraídos de *L. interrogans* sorovar Copenhageni Fiocruz L1-130.

MLST 1	MLST 2	MLST 3	Fiocruz L1-130
glmU	*adk*	*adk*	*rrs*
pntA	*glmU*	*icdA*	*secY*
sucA	*icdA*	*lipL32*	*rpoB*
tpiA	*lipL32*	*lipL41*	*ligB*
pfkB	*lipL41*	*rrs2*	*lipL41*
mreA	*mreA*	*secY*	
caiB	*pntA*		

Tabela 2. Informações sobre tamanho (em pares de base), posição no genoma e LOCUS_TAG de cada gene utilizado neste trabalho.

Gene	Tamanho (pb)	Posição no genoma	LOCUS_TAG
glmU	759	489012 - 489770	LIC_10428
pntA	1170	56259 - 57428	LIC_10046
sucA	921	2995070 - 2997832	LIC_12474
tpiA	753	2533875 - 2534627	LIC_12094
pfkB	329	2840289 - 2841275	LIC_12346
mreA	508	1560465 - 1558942	LIC_11262
caiB	391	1168252 - 1169424	LIC_10969
adk	564	3458298 - 3458861	LIC_12852
icdA	1197	3979829 - 3981025	LIC_13244
lipL32	819	1666299 - 1667117	LIC_11352
lipL41	1068	3603575 - 3604642	LIC_12966
rrs2	1512	1862433 - 1863944	LIC_11508
secY	1383	3458869 - 3460251	LIC_12853
rpoB	3681	912196 - 915876	LIC_10753
ligB	9673	536395 - 532067	LIC_10464

Todos os 600 *drafts* de *Leptospira* (aproximadamente 2 bilhões e 700 mil pares de base disponibilizados em formato digital) foram recuperadas do GenBank usando *scripts* escritos em linguagem Python através das palavras-chave "*Leptospira* [organism]", sendo a busca realizada através do módulo Entrez do pacote BioPython. A distribuição das espécies de *Leptospira* estão apresentadas na Tabela 3. As sequências de cada isolado foram indexadas em um banco de dados com base na anotação quanto à sua espécie, sorogrupo, sorovar e cepa.

Tabela 3. Distribuição das sequências genômicas de espécies do gênero *Leptospira* recuperadas do GenBank.

Espécie	Número de cepas
L. interrogans	391
L. kirschneri	50
L. santarosai	47
L. borgpetersenii	40
L. noguchii	18
L. weilii	18
L. licerasiae	8
L. alstonii	4
L. meyeri	4
L. kmetyi	2
L. broomii	2
L. alexanderi	2
L. inadai	2
L. wolffii	2
L. fainei	2
L. terpstrae	2
L. wolbachii	2
L. yanagawae	2
L. vanthielii	2
Total	600

A busca das sequências homólogas dos 25 *loci* foi realizada em cada um dos isolados indexados no nosso banco de dados através do BLASTn do pacote NCBI-BLAST+. O tamanho de cada sequência homólogas obtida no BLAST foi padronizado de acordo com a sequência de maior tamanho. Nos casos de genes fragmentados (presentes nas extremidades de *scaffolds* de genomas *drafts* ou sequências parciais de genes), as regiões não alinhadas foram preenchidas com gaps ('N'). Os isolados que não apresentaram sequências homólogas de determinados *loci* tiveram os campos da tabela preenchidos com sinais de '?', sendo este passo também seguido para as regiões de genes encontrados em extremidades de *scaffolds*, indicando que a região pode ser desconhecida. Em casos de gaps de alinhamento, as regiões faltantes foram indicadas com '-'.

As sequências homólogas de cada *loci* (referente a cada isolado) foram alinhadas com auxílio do MUSCLE (*multiple sequence comparison by log-expectation*), um software para alinhamento múltiplo de sequências, e gerou um alinhamento múltiplo em formato FASTA. O conjunto de dados final foi constituído pelas sequências homólogas de cada lócus (obtido no alinhamento múltiplo) referente a cada um dos 600 genomas e estruturado usando o formato de arquivo

CSV. Às classes e atributos foram atribuídos nomes das cepas e a sequência de nucleotídeos homólogos referentes a cada cepa, respectivamente. As classes foram dispostas na primeira coluna do arquivo CSV e, nas demais, os atributos (Tabela 3).

Tabela 4. Exemplo da disposição dos dados genômicos no arquivo CSV.

		MLST 1	
	glmU	**pntA**	...
Genoma 1	Homólogo 1	Homólogo 1	...
Genoma 2	Homólogo 2	Homólogo 2	...
.	...	...	...
.	...	...	...
.	...	...	...
Genoma 600	Homólogo 600	Homólogo 600	...

Na primeira coluna foram alocadas as classes (nome das cepas) e nas demais, os atributos (sequência homóloga referente a cada cepa). A tabela mostra apenas os *loci glmU* e *pntA* referentes ao MLST 1.

4.2 Aplicação dos algoritmos de mineração

Para a análise dos polimorfismos e posterior classificação dos isolados de *Leptospira* foram utilizados três algoritmos computacionais de classificação: C4.5, Naive Bayes e SVM, todos implementados no Weka (*Waikato Environment for Knowledge Analysis*), um *software* gratuito e de código aberto que possui um conjunto de algoritmos de aprendizado de máquina para tarefas de mineração de dados (Hall et al., 2009). O Weka possui uma interface gráfica (conhecida como Weka Explorer), sendo uma das mais populares ferramentas de mineração de dados utilizada por pesquisadores em ambiente acadêmico. O banco de dados foi avaliado no software Weka utilizando os parâmetros *default* e com um *fold* de 10%.

4.3 Avaliação das metodologias de classificação

O poder de classificação de cada algoritmo foi avaliado através da comparação das acurácias obtidas em cada um deles. O banco de dados utilizado como entrada para os três algoritmos computacionais foi também submetido ao método MLST (esquemas 1, 2 e 3), para isso, diferentes combinações de subconjuntos de genes foram testados. O processo iniciou com o teste de quatro genes do esquema de MLST 1, quatro genes do esquema de MLST 2 e quatro

genes do esquema de MLST 3 e a cada etapa adicionava um gene, e assim sucessivamente totalizando 12 testes. Ambas as metodologias avaliaram os polimorfismos recuperados dos 600 genomas.

5 RESULTADOS

5.1 Aplicação dos algoritmos de mineração de dados

A classificação dos isolados de *Leptospira* foi conduzida usando os algoritmos de mineração de dados C4.5, Naive Bayes e SVM implementados no *software* Weka. A análise realizada por cada algoritmo gerou matrizes de confusão, acurácias de classificação (média das acurácias de cada classe), e uma árvore de decisão, gerada apenas pelo algoritmo C4.5. As matrizes de confusão são utilizadas para avaliar o desempenho de classificação de cada algoritmo e mostram o número de classificações analisadas correta (na diagonal principal) e incorretamente para cada espécie. As outras frequências representam os erros de classificação.

Para C4.5, por exemplo, de 600 *drafts* de *Leptospira* analisadas, 574 foram classificadas corretamente, obtendo uma acurácia de 95,6% (Tabela 5). Das 26 cepas classificadas incorretamente, 3 cepas de *L. interrogans* foram classificadas como *L. noguchii*, 2 cepas *L. noguchii* e 2 cepas de *L. wolffii* foram classificadas como *L. interrogans*, 1 cepa de *L. licerasiae* foi classificada como *L. weilii*, 2 cepas de *L. kmetyi* como *L. borgpetersenii*, 2 cepas de *L. yanagawae* como *L. licerasiae*, 2 cepas de *L. alexanderi* como *L. alstonii*, 2 cepas de *L. fainei* como *L. inadai*, 2 cepas de *L. santarosai* como *L. kirschneri*, 1 cepa de *L. inadai* como *L. fainei*, e, por fim, 2 cepas de *L. terpstrae*, 2 cepas de *L. wolbachii* e 2 cepas de *L. vanthielii* como *L. meyeri*.

Tabela 5. Matriz de confusão gerada pelo algoritmo C4.5.

Classificado como →	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s
L. noguchii = a	**16**	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L. interrogans = b	3	**388**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L. weilii = c	0	0	**18**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L. santarosai = d	0	0	0	**45**	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
L. borgpeters. = e	0	0	0	0	**40**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L. alexanderi = f	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0
L. kmetyi = g	0	0	0	0	2	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0
L. licerasiae = h	0	0	1	0	0	0	0	**7**	0	0	0	0	0	0	0	0	0	0	0
L. alstonii = i	0	0	0	0	0	0	0	0	**4**	0	0	0	0	0	0	0	0	0	0
L. inadai = j	0	0	0	0	0	0	0	0	0	**0**	0	0	1	1	0	0	0	0	0
L. kirschneri = k	0	0	0	0	0	0	0	0	0	0	**50**	0	0	0	0	0	0	0	0
L. wolffii = l	0	2	2	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0
L. broomii = m	0	0	0	0	0	0	0	0	0	0	0	0	**2**	0	0	0	0	0	0
L. fainei = n	0	0	0	0	0	0	0	0	0	2	0	0	0	**0**	0	0	0	0	0
L. terpstrae = o	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	2	0	0
L. wolbachii = p	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	2	0	0
L. meyeri = q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**4**	0	0
L. yanagawae = r	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	**0**	0
L. vanthielii = s	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	**0**

O algoritmo Naive Bayes também obteve um alto desempenho, pois, assim como C4.5, classificou corretamente 574 cepas e apresentou uma acurácia de 95,6% (Tabela 6). Para Naive Bayes, as cepas classificadas incorretamente foram: 4 cepas de *L. interrogans* classificadas como *L. noguchii*, 2 cepas *L. alexanderi* e 1 cepa de *L. kmetyi* classificadas como *L. weilii*, 1 cepa de *L. wolffii* e 1 cepa de *L. broomii* como *L. santarosai*, 2 cepas de *L. inadai* como *L. borgpetersenii*, 1 cepa de *L. wolffii*, 1 cepa de *L. broomii* e 1 cepa de *L. fainei* como *L. licerasiae*, 1 cepa de *L. fainei* como *L. broomii*, 2 cepas de *L. terpstrae*, 2 cepas de *L. wolbachii*, 2 cepas de *L. yanagawae* e 2 cepas de *L. vanthielii* classificadas como *L. meyeri*.

Tabela 6. Matriz de confusão gerada pelo algoritmo Naive Bayes.

Classificado como →	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s
L. noguchii = a	**18**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L. interrogans = b	4	**387**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L. weilii = c	0	0	**18**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L. santarosai = d	0	0	0	**44**	0	0	1	0	0	0	2	0	0	0	0	0	0	0	0
L. borgpeters. = e	0	0	0	0	**40**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L. alexanderi = f	0	0	2	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0
L. kmetyi = g	0	0	1	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0
L. licerasiae = h	0	0	0	0	0	0	0	**8**	0	0	0	0	0	0	0	0	0	0	0
L. alstonii = i	0	0	0	0	0	0	0	0	**4**	0	0	0	0	0	0	0	0	0	0
L. inadai = j	0	0	0	0	2	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0
L. kirschneri = k	0	0	0	0	0	0	0	0	0	0	**50**	0	0	0	0	0	0	0	0
L. wolffii = l	0	0	0	1	0	0	0	1	0	0	0	**0**	0	0	0	0	0	0	0
L. broomii = m	0	0	0	1	0	0	0	1	0	0	0	0	**1**	0	0	0	0	0	0
L. fainei = n	0	0	0	0	0	0	0	1	0	0	0	0	0	**0**	0	0	0	0	0
L. terpstrae = o	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	2	0	0
L. wolbachii = p	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	2	0	0
L. meyeri = q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**4**	0	0
L. yanagawae = r	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	2	**0**	0
L. vanthielii = s	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	**0**

Por fim, dentre 600 *drafts* de *Leptospira*, o algoritmo de mineração de dados SVM conseguiu classificar corretamente 559 e atingiu uma acurácia de 93,2%. Um desempenho não tão bom quando comparado aos dois primeiros algoritmos, mas ainda assim um alto desempenho. As cepas classificadas incorretamente foram: 3 cepas de *L. interrogans* classificadas como *L. noguchii*, 8 cepas de *L. licerasiae*, 2 cepas de *L. inadai*, 2 cepas de *L. wolffii*, 2 cepas de *L. broomii*, 2 cepas de *L. fainei*, 2 cepas de *L. terpstrae*, 2 cepas de *L. wolbachii*, 4 cepas de *L. meyeri*, 2 cepas de *L. yanagawae* e 2 cepas de *L. vanthielii* classificadas como *L. interrogans*, 1 cepa de *L. kmetyi*, 4 cepas de *L. alstonii* como *L. weilii*, 1 cepa de *L. kmetyi* como *L. santarosai*, 2 cepas de *L. alexanderi* como *L. borgpetersenii* (Tabela 7).

Tabela 7. Matriz de confusão gerada pelo algoritmo SVM.

Classificado como →	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s
L. noguchii = a	**18**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L. interrogans = b	3	**388**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L. weilii = c	0	0	**18**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L. santarosai = d	0	0	0	**45**	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
L. borgpeters. = e	0	0	0	0	**40**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L. alexanderi = f	0	0	1	0	2	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0
L. kmetyi = g	0	0	0	1	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0
L. licerasiae = h	0	8	4	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0
L. alstonii = i	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0
L. inadai = j	0	2	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0
L. kirschneri = k	0	0	0	0	0	0	0	0	0	0	**50**	0	0	0	0	0	0	0	0
L. wolffii = l	0	2	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0
L. broomii = m	0	2	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0
L. fainei = n	0	2	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0
L. terpstrae = o	0	2	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0
L. wolbachii = p	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0
L. meyeri = q	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0
L. yanagawae = r	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0
L. vanthielii = s	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**

A Figura 1 mostra a árvore de decisão gerada somente pelo algoritmo C4.5 para classificação de espécies de *Leptospira*. A árvore de decisão é formada por um conjunto de raíz, ramos, nós e folhas que permitem a vizualização de cada caso. A árvore inicia a classificação (raíz) com o gene icdA, os ramos mostram os nucleotídeo encontrados na posição 69 do gene icdA, cada nó representa o próximo teste realizado nos atributos e cada folha representa a classe prevista para o atributo testado (Figura 1).

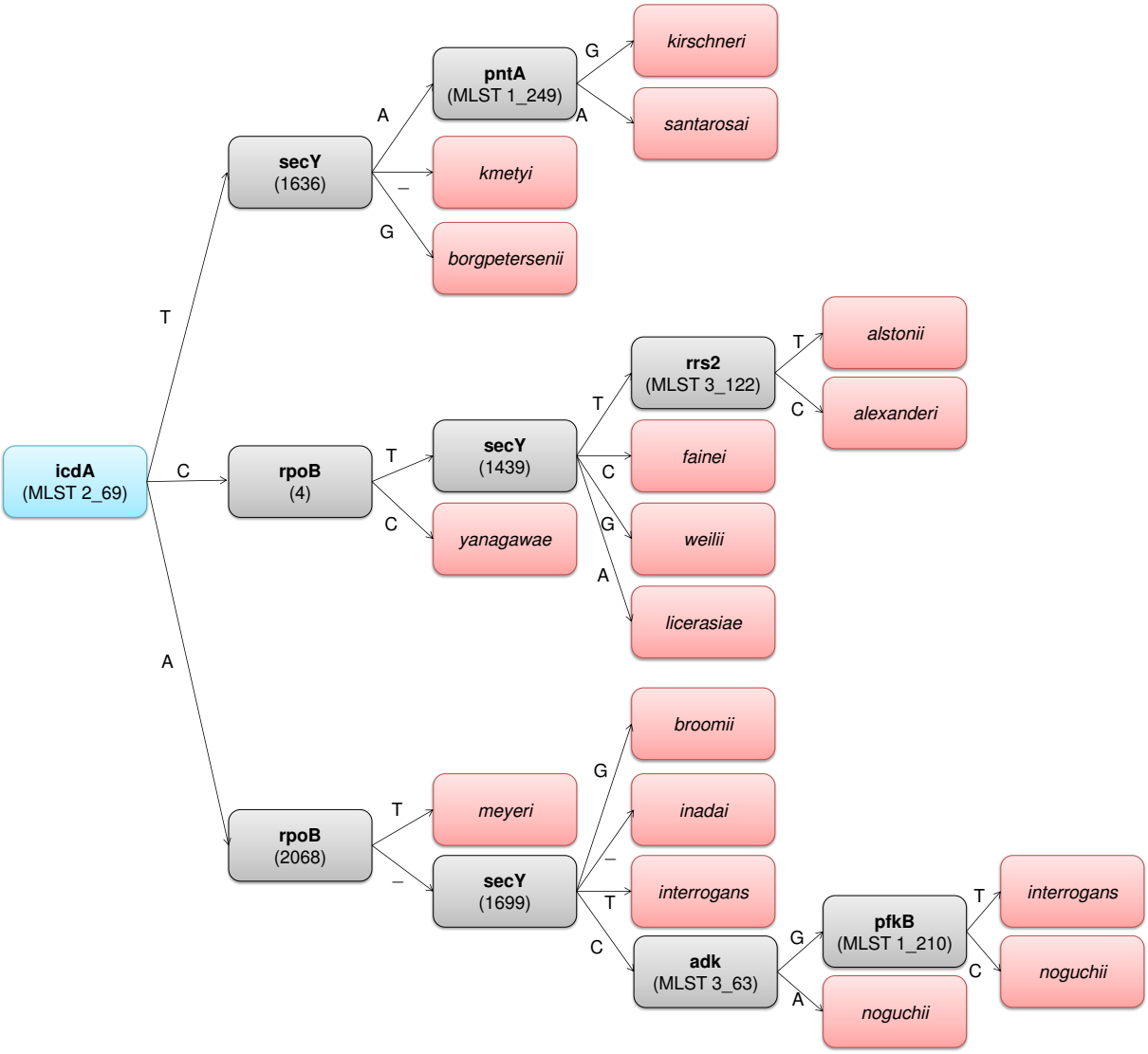

Figura 1. Árvore de decisão gerada pelo algoritmo C4.5 para classificar espécies de Leptospira. A raíz astá representada em azul, os ramos estão representados pelos nucleotídeos, os nós estão representados em cinza e as folhas estão representadas em rosa.

Dentre as 19 espécies de *Leptospira* presentes no nosso banco de dados (Tabela 2), o algoritmo J48 classificou com precisão *L. noguchii*, *L. interrogans*, *L. broomii*, *L. inadai*, *L. meyeri*, *L. licerasiae*, *L. weilii*, *L. fainei*, *L. alstonii*, *L. alexanderi*, *L. yanagawae*, *L. santarosai*, *L. kirschneri*, *L. borgpetersenii* e *L. kmetyi*, totalizando 15 espécies (quadros cor de rosa). Para isso, analisou padrões de nucleotídeos encontrados em apenas 7 genes, sendo eles: MLST 1 = *pfkB* e *pntA*, MLST 2 = *icdA*, MLST 3 = *adk* e *rrs2*, Fiocruz L1-130 = *rpoB* e *secY* (quadros de cor cinza).

5.2 Avaliação das metodologias de classificação

A Figura 2 apresenta as acurácias obtidas a partir da análise de mineração de dados realizadas pelos três algoritmos computacionais – C4.5, Naive Bayes e SVM – e as acurácias obtidas nos testes com a metodologia MLST.

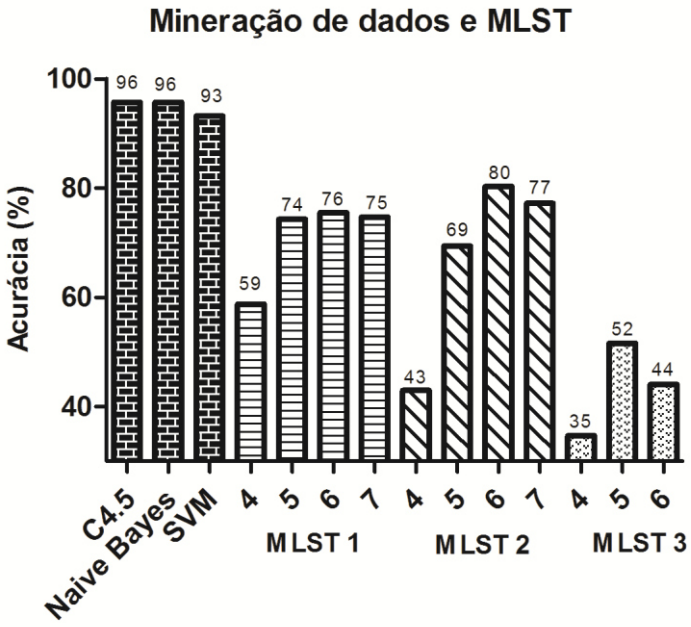

Figura 2. Classificação de Leptospira spp. predita por algoritmos computacionais baseados em mineração de dados e pela metodologia do MLST. As acurácias foram obtidas a partir da análise de mineração de dados realizada pelos três algoritmos computacionais - C4.5, Naive Bayes e SVM - e a partir dos testes com a metodologia MLST. As três primeiras barras referem-se as acurácias obtidas pelos algoritmos e as demais referem-se as acurácias dos testes para cada esquema MLST. Os números 4, 5, 6 e 7 (Esquema 1 e 2) e 4, 5 e 6 (Esquema 3) referem-se ao número mínimo de genes necessários para iniciar uma tentativa de classificação pela metodologia MLST.

As três primeiras barras referem-se aos algoritmos de mineração de dados e as demais se referem aos testes realizados para cada metodologia MLST. Os números 4, 5, 6 e 7 (esquema 1 e 2) e os números 4, 5 e 6 (esquema 3) referem-se ao número mínimo de genes exigidos pela metodologia de MLST para iniciar uma tentativa de classificação.

6 DISCUSSÃO GERAL E PERSPECTIVAS

6.1 Comparação dos resultados gerados pelos algoritmos de mineração de dados

Neste trabalho, 600 sequências *draft* de *Leptosira* pertencentes a 19 espécies diferentes foram avaliadas através da aplicação de três algoritmos de mineração de dados - C4.5, Naive Bayes e SVM. As matrizes de confusão geradas pelos três algoritmos apresentaram as cepas classificadas correta e incorretamente. C4.5 e Naive Bayes apresentaram 26 cepas classificadas incorretamente enquanto SVM apresentou 41. Esses erros na classificação ocorrem devido à pouca informação genética disponível para cada espécie. Por exemplo, para *L. alexanderi*, *L. kmetyi*, *L. inadai*, *L. wolffii*, *L. broomii*, *L. fainei*, *L. terpstrae*, *L. wolbachii*, *L. yanagawae* e *L. vanthielii*, somente 2 sequências genômicas estão disponíveis para cada uma delas nos bancos de dados, ou seja, apenas 2 genomas foram utilizados na análise de mineração (Tabela 1). Como algoritmos de mineração de dados atuam buscando padrões entre as sequências genéticas, é necessário que mais sequências da mesma espécie estejam disponíveis, de forma que os algoritmos possam trabalhar com mais informações sobre a mesma sorovar, ou seja, com mais repetitividade. Além da matriz de confusão, o algoritmo C4.5 gerou uma árvore de decisão (Figura 1) onde mostra as espécies que foram classificadas por este algoritmo. Das 19 espécies presentes no nosso banco de dados, C4.5 conseguiu classificar 15 delas com uma acurácia de aproximadamente 96%. A acurácia desta classificação pode ser vista na matriz de confusão gerada por este algoritmo (Tabela 5). As árvores de decisão possuem uma representação intuitiva, que, em geral, torna o modelo de classificação fácil de ser interpretado. Além da árvore de decisão gerada pelo algoritmo C4.5 ser facilmente compreendida, ela apresenta a contribuição de cada atributo para o resultado final da análise, ou seja, observando a árvore, é possível saber que, de um total de 25 *loci* avaliados na análise de classificação, apenas os nucleotídeos polimórficos de 7 *loci* foram utilizados para classificar as 15 espécies de *Leptospira*. Este é um resultado bastante positivo, pois, além de C4.5 apresentar uma alta acurácia, ele nos mostra quais os genes contribuíram para a classificação e revela as posições onde se encontram os nucleotídeos polimórficos dentro de cada *loci*. Naive Bayes e SVM apresentaram

uma acurácia de classificação de 95.6% e 93%, respectivamente. O desempenho de SVM foi menos satisfatório quando comparado aos dois primeiros algoritmos, mas ainda assim, um bom desempenho. Ainda assim, Naive Bayes e SVM possuem algumas desvantagens quando comparados com C4.5: além de consumirem mais tempo para apresentarem os resultados (C4.5 = 2 a 3h, Naive Bayes = 24h e SVM = 4h), ambos não geram gráficos de representação, de forma que não é possível saber quais atributos contribuíram na análise de classificação. Desta forma, pode-se dizer que, entre os 3 algoritmos utilizados, C4.5 é o mais adequado para ser usado na classificação de espécies de *Leptospira*, pois, além de apresentar os atributos que contribuíram para análise e alcançar a melhor acurácia, este algoritmo consumiu menos tempo para realizar o trabalho de classificação.

6.2 Relação entre a metodologia de mineração de dados e a metodologia MLST

A figura 2 mostra o resultado final deste trabalho onde apresenta as acurácias das duas metodologias de classificação de *Leptospira*, a primeira utilizando algoritmos de mineração de dados e a segunda utilizando a metodologia MLST. Estes resultados foram gerados a partir das informações presentes no nosso banco de dados. A figura mostra uma grande discordância entre as acurácias preditas pelos algoritmos computacionais e as acurácias preditas por cada metodologia MLST. Enquanto os três algoritmos atingiram acurácias acima de 93%, as metodologias MLST atingiram, no máximo, 80% (MLST 2 testado com 6 genes). Esta discordância deve-se ao fato de que o banco de dados submetido aos algoritmos de mineração e também aos 3 esquemas de MLST possui informações genômicas de 19 espécies, e cada metodologia MLST foi desenvolvida para classificar um número bem reduzido de espécies. Por exemplo, o MLST 1 é utilizado para classificar 7 espécies (Thaipadungpanit et al., 2007; Boonsilp et al., 2013) e o MLST 2 e MLST 3 são utilizados para classificar apenas 2 espécies (Varni et al., 2014; Ahmed et al., 2006).

6.3 Relação entre a mineração de dados e o futuro da classificação de espécies do gênero *Leptospira*

Um grande número de abordagens moleculares baseadas na análise da sequência nucleotídica tem sido utilizado para identificação de espécies (Ahmed, & P. Grobusch, 2012). Por exemplo, para classificação de espécies de *Leptospira* pode-se usar a análise da sequência dos genes *rrs* (Morey et al., 2006), *rpoB* (Scola et al., 2006) e *gyrB* (Slack et al., 2006). Agora, para a classificação de isolados de *Leptospira* a nível de sorovar através de mineração de dados, um conjunto maior de informações genômicas precisam estar disponíveis. Sendo assim, a mineração de dados poderá auxiliar no conciliamento entre identidade sorológica e molecular. Recentemente, pesquisadores anunciaram que alguns STs foram encontrados em sorovares que pertenciam ao mesmo sorogrupo (Varni et al., 2014; Goarant, 2014) demonstrando que esta associação pode ser possível.

Com a redução no custo do sequenciamento (Buermans & den Dunnen, 2014), o número de genomas sequenciados de *Leptospira* aumentará substancialmente, e cada vez mais informações genômicas deste gênero estarão disponíveis nos bancos de dados públicos. Com isso, a aplicação de algoritmos baseado em mineração de dados poderá ser realizado com dados cada vez mais robustos e com acurácias cada vez mais próximas de 100%. Além disso, associações entre a sequência nucleotídica do patógeno e animais hospedeiros ou nicho ambiental também poderão ser previstas e decisões sobre estratégias de controle e prevenção de doenças poderão ser discutidas antecipadamente.

7 CONCLUSÃO GERAL

Neste estudo, nós relatamos pela primeira vez a utilização de algoritmos de mineração de dados para fins de estudo de classificação de cepas do gênero *Leptospira*. Foram utilizados os algoritmos C4.5, Naive Bayes e SVM e todos apresentaram acurácias acima de 93% frente a um banco de dados composto por 600 isolados pertencentes a 19 espécies de *Leptospira*. Dentre os três algoritmos aplicados, pode-se concluir que C4.5 é o mais adequado para classificar espécies deste gênero, pois, além de mostrar quais *loci* contribuíram na análise, consumiu menos tempo e alcançou o melhor valor de acurácia quando comparado aos outros algoritmos. Com a redução no custo do sequenciamento de DNA e o consequente aumento de informações genômicas disponíveis, esta metodologia poderá ser realizada com dados cada vez mais robustos, de forma que será possível classificar isolados de *Leptospira* não só em nível de espécie, mas também em nível de sorovar. Sendo assim, a mineração de dados poderá traçar uma correlação unidirecional entre perfil alélico e sorovar. Além disso, esta tecnologia pode ser utilizada não só para classificação de *Leptospira*, mas também para classificação de qualquer outro organismo procariótico e até eucariótico, desde que os genes classificadores estejam bem estabelecidos para análise.

8 REFERÊNCIAS

ADLER, B. (2014). Leptospira and Leptospirosis. *Current Topics in Microbiology and Immunology*. 387. p. 293.

AHMED, N., DEVI, S.M., VALVERDE, M. DE LOS A., VIJAYACHARI, P., MACHANG'U, R.S., ELLIS, W.A. & HARTSKEERL, R.A. (2006). Multilocus sequence typing method for identification and genotypic classification of pathogenic Leptospira species. *Annals of clinical microbiology and antimicrobials*. 5. p. 1–10.

AHMED, A. & P. GROBUSCH, M. (2012). Molecular Approaches in the Detection and Characterization of Leptospira. *Journal of Bacteriology & Parasitology*. 03(02).

BANERJEE, A.K., RAVI, V., MURTY, U.S.N., SENGUPTA, N. & KARUNA, B. (2013). Application of intelligent techniques for classification of bacteria using protein sequence-derived features. *Applied biochemistry and biotechnology*. 170(6). p. 1263–81.

BELÉN, A., PAVÓN, I. & MAIDEN, M.C.J. (2009). Multilocus Sequence Typing. *Methods in Molecular Biology*. 551(8). p. 1–11.

BHARTI, A.R., NALLY, J.E., RICALDI, J.N., MATTHIAS, M.A., DIAZ, M.M., LOVETT, M.A., LEVETT, P.N., GILMAN, R.H., WILLIG, M.R., GOTUZZO, E. & VINETZ, J.M. (2003). Reviews Leptospirosis : a zoonotic disease of global importance. *The Lancet Infectious Diseases*. p. 757–771.

BOONSILP, S., THAIPADUNGPANIT, J., AMORNCHAI, P., WUTHIEKANUN, V., BAILEY, M.S., HOLDEN, M.T.G., ZHANG, C., JIANG, X., KOIZUMI, N., TAYLOR, K., GALLOWAY, R., HOFFMASTER, A.R., CRAIG, S., SMYTHE, L.D., HARTSKEERL, R.A., DAY, N.P., CHANTRATITA, N., FEIL, E.J., AANENSEN, D.M., ET AL. (2013). A single multilocus sequence typing (MLST) scheme for seven pathogenic Leptospira species. *PLoS neglected tropical diseases*. 7(1). p. 1–10.

BUERMANS, H.P.J. & DEN DUNNEN, J.T. (2014). Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*. 1842(10). p. 1932–1941.

CERQUEIRA, G.M., MCBRIDE, A.J.A., HARTSKEERL, R.A., AHMED, N., DELLAGOSTIN, O.A., ESLABÃO, M.R. & NASCIMENTO, A.L.T.O. (2010). Bioinformatics describes novel Loci for high resolution discrimination of leptospira isolates. *PloS one*. 5(10). p. E15335.

CHEN, S.Y. & LIU, X. (2009). The contribution of data mining to information science. *Journal of Information Science*. 30(6). p. 550–558.

CORTES, C. & VAPNIK, V. (1995). Support-Vector Networks. *Machine Learning*. 20. p. 273–297.

GALLOWAY, R.L. & LEVETT, P.N. (2010). Application and validation of PFGE for serovar identification of Leptospira clinical isolates. *PLoS neglected tropical diseases*. 4(9).

GOARANT, C. (2014). Leptospirosis: Time to move to molecular epidemiology. Comments on "Reassessment of MLST schemes for Leptospira spp. typing worldwide" by Varni and colleagues. *Infection, Genetics and Evolution*.

21(JANUARY 2014). p. 484–485.

HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P. & WITTEN, I.H. (2009). The WEKA data mining software: An Update. *SIGKDD Explorations*. 11(1). p. 10.

HAN, J., KAMBER, M. & PEI, J. (2012). *Data Mining: Concepts and Techniques* 3rd editio ed., Waltham, USA: Elsevier.

HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. (2009). *The Elements of Statistical Learning* Second edi ed. Springer (ed.)., New York.

JOLLEY, K. A, CHAN, M.-S. & MAIDEN, M.C.J. (2004). mlstdbNet - distributed multi-locus sequence typing (MLST) databases. *BMC bioinformatics*. 5. p. 86.

KMETY, E. & DIKKEN, H. (1993). *Classification of the species Leptospira interrogans and history of its serovars*, Groningen-Netherlands.

LEVETT, P.N. (2015). Leptospira and Leptospirosis. *Current Topics in Microbiology and Immunology*. 387. p. 11–20.

LEVETT, P.N. (2007). Sequence-based typing of leptospira: epidemiology in the genomic era. *PLoS neglected tropical diseases*. 1(2). p. 1–2.

LEVETT, P.N., MOREY, R.E., GALLOWAY, R.L. & STEIGERWALT, A.G. (2006). Leptospira broomii sp. nov., isolated from humans with leptospirosis. *International journal of systematic and evolutionary microbiology*. 56(Pt 3). p. 671–3.

MAIDEN, M.C., BYGRAVES, J. A, FEIL, E., MORELLI, G., RUSSELL, J.E., URWIN, R., ZHANG, Q., ZHOU, J., ZURTH, K., CAUGANT, D. A, FEAVERS, I.M., ACHTMAN, M. & SPRATT, B.G. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences of the United States of America*. 95(March). p. 3140–3145.

MAIDEN, M.C.J. (2006). Multilocus sequence typing of bacteria. *Annual review of microbiology*. 60. p. 561–588.

MAJED, Z., BELLENGER, E., POSTIC, D., POURCEL, C., BARANTON, G. & PICARDEAU, M. (2005). Identification of Variable-Number Tandem-Repeat Loci in Leptospira interrogans Sensu Stricto. *Journal of Clinical Microbiology*. 43(2). p. 539–545.

MOREY, R.E., GALLOWAY, R.L., BRAGG, S.L., STEIGERWALT, A.G., MAYER, L.W. & LEVETT, P.N. (2006). Species-specific identification of Leptospiraceae by 16S rRNA gene sequencing. *Journal of Clinical Microbiology*. 44(10). p. 3510–3516.

NALAM, K., AHMED, A., DEVI, S.M., FRANCALACCI, P., BAIG, M., SECHI, L.A., HARTSKEERL, R.A. & AHMED, N. (2010). Genetic Affinities within a Large Global Collection of Pathogenic Leptospira: Implications for Strain Identification and Molecular Epidemiology. *PLoS ONE*. 5(8). p. E12637.

POSTIC, D., RIQUELME-SERTOUR, N., MERIEN, F. & PEROLAT, P. (2000). Interest of partial 16S rDNA gene sequences to resolve heterogeneities between Leptospira collections : application to L . meyeri. *Research in Microbiology*. 151. p. 333–341.

PRANEENARARAT, T., TAKAGI, T. & IWASAKI, W. (2012). Integration of interactive, multi-scale network navigation approach with Cytoscape for functional genomics in the big data era. *BMC genomics*. 13. p. 1–10.

ROMERO, E.C. & YASUDA, P.H. (2006). Molecular characterization of Leptospira sp . strains isolated from human subjects in São Paulo , Brazil using a polymerase chain reaction-based assay : a public health tool. *Memórias do instituto Oswaldo Cruz*. 101(June). p. 373–378.

RUSSELL, S. & NORVIG, P. (1995). *Artificial Intelligence: A Modern Approach*, New Jersey: Prentice Hall.

SALZBERG, S. (1994). Book Review: C4.5: Programs for Machine Learning. *Machine Learning*. 1(16). p. 235–240.

SCOLA, B. LA, BUI, L.T.M., BARANTON, G., KHAMIS, A. & RAOULT, D. (2006). Partial rpoB gene sequencing for identification of Leptospira species. *FEMS Microbiology Letters*. 263(Table 1). p. 142–147.

SHAH, S.C. & KUSIAK, A. (2004). Data mining and genetic algorithm based gene / SNP selection. *Artificial Intelligence in Medicine*. p. 183–196.

SLACK, A.T., SYMONDS, M.L., DOHNT, M.F. & SMYTHE, L.D. (2006). Identification of pathogenic Leptospira species by conventional or real-time PCR and sequencing of the DNA gyrase subunit B encoding gene. *BMC microbiology*. 6. p. 95.

THAIPADUNGPANIT, J., WUTHIEKANUN, V., CHIERAKUL, W., SMYTHE, L.D., PETKANCHANAPONG, W., LIMPAIBOON, R., APIWATANAPORN, A., SLACK, A.T., SUPUTTAMONGKOL, Y., WHITE, N.J., FEIL, E.J., DAY, N.P.J. & PEACOCK, S.J. (2007). A dominant clone of Leptospira interrogans associated with an outbreak of human leptospirosis in Thailand. *PLoS Neglected Tropical Diseases*. 1(1). p. 1–6.

VARNI, V., RUYBAL, P., LAUTHIER, J.J., TOMASINI, N., BRIHUEGA, B., KOVAL, A. & CAIMI, K. (2014). Reassessment of MLST schemes for Leptospira spp. typing worldwide. *Infection, Genetics and Evolution*. 22. p. 216–222.

VIJAYACHARI, P., AHMED, N., SUGUNAN, A. P.P., GHOUSUNNISSA, S., RAO, K.R.R., HASNAIN, S.E.E. & SEHGAL, S.C. (2004). Use of fluorescent amplified fragment length polymorphism for molecular epidemiology of leptospirosis in India. *Journal of Clinical Microbiology*. 42(8). p. 3575–3580.

ZHUANG, Z.Y., CHURILOV, L., BURSTEIN, F. & SIKARIS, K. (2009). Combining data mining and case-based reasoning for intelligent decision support for pathology ordering by general practitioners. *European Journal of Operational Research*. 195(3). p. 662–675.