OTIMIZAÇÃO DE TÉCNICAS DE CLASSIFICAÇÃO DE CEPAS DE *Leptospira* SPP. ATRAVÉS DE MÉTODOS COMPUTACIONAIS BASEADOS EM MINERAÇÃO DE DADOS

JULIA LABONDE[1]; MARCUS REDÜ ESLABÃO[2]; FREDERICO SCHMITT KREMER[3]; MONIZE PROVISOR[4]; ODIR ANTONIO DELLAGOSTIN[5]

[1]*Laboratório de Proteômica e Bioinformática, Núcleo de Biotecnologia – CDTec, UFPel - julialabonde@gmail.com;*
[2]*Laboratório de Proteômica e Bioinformática, Núcleo de Biotecnologia – CDTec, UFPel - marcus.eslabao@yahoo.com.br;*
[3]*Laboratório de Proteômica e Bioinformática, Núcleo de Biotecnologia – CDTec, UFPel - fred.s.kremer@gmail.com;*
[4]*Laboratório de Proteômica e Bioinformática, Núcleo de Biotecnologia – CDTec, UFPel - moprovisor@hotmail.com;*
[5] *Núcleo de Biotecnologia – CDTec, UFPel - odir@ufpel.edu.br*

1. INTRODUÇÃO

A leptospirose é uma doença infecciosa de importância mundial que afeta humanos e animais, causada por espiroquetas patogênicas pertencentes ao gênero *Leptospira*. Os roedores são os principais hospedeiros das espiroquetas, as quais são eliminadas no ambiente através de sua urina. A infecção humana resulta do contato direto ou indireto com a urina contaminada de animais portadores (ADLER; MOCTEZUMA, 2010). A doença tem ampla distribuição geográfica, ocorrendo em áreas de clima temperado e tropical. Dados da Organização Mundial da Saúde revelam que mais de 500.000 casos de leptospirose ocorrem a cada ano em todo o mundo. Por possuir diagnóstico ineficiente e ser facilmente confundida com outras doenças (febre amarela, gripe ou dengue), a leptospirose ainda é uma doença negligenciada, sendo o seu número de casos subestimado (WHO, 2011).

Para a área epidemiológica e sua consequente tomada de decisão com relação à saúde pública, é fundamental que os laboratórios tenham a capacidade de identificar com precisão as cepas de *Leptospira* spp. que causam doença (AHMED et al., 2006). A principal metodologia de identificação das cepas é o teste de aglutinação microscópica (MAT). No entanto, este teste é demorado e exige a presença de uma coleção completa de cepas do gênero *Leptospira*, tornando esta metodologia cara e restrita a laboratórios de referência (ROMERO; YASUDA, 2006).

O *Multilocus sequence typing* (MLST) é um método de genotipagem baseado na análise da sequência de genes *housekeeping,* e tem por objetivo proporcionar um sistema de tipagem discriminativo e portátil para bactérias e outros organismos (MAIDEN et al., 1998). Para o gênero *Leptospira* já existem bancos de dados *online* contendo sequências desses genes (THAIPADUNGPANIT et al., 2007). Outros genes como *rrs*, *rpoB*, *SecY*, *LigB* e *LipL41* (ausentes no banco de dados do MLST) também estão sendo apontados como promissores para a classificação (CERQUEIRA et al., 2010; MOREY et al., 2006; SCOLA et al., 2006).

As sequências polimórficas desses genes de *Leptospira* são importantes quando se objetiva classificar diferentes isolados patogênicos, no entanto, a quantidade de dados genômicos disponíveis nesses bancos de dados é incomensurável. Desta forma, algoritmos de mineração de dados são necessários para minerar as informações existentes. Algoritmos computacionais baseados em

árvores de decisão, redes neurais, Naive Bayes e SVM já são utilizados para manipular dados genômicos, e podem utilizar as sequências gênicas disponibilizadas nos bancos de dados do MLST e também as sequências dos outros genes citados para classificar *Leptospira* de forma mais precisa e acurada.

Sendo assim, o objetivo deste trabalho foi realizar uma análise de mineração de dados utilizando algoritmos computacionais com base nas sequências gênicas e genômicas de *Leptospira* spp., disponíveis no GenBank, com o intuito de buscar polimorfismos que nos forneçam informações precisas e acuradas para uma correta identificação de isolados.

2. METODOLOGIA

Criação do *dataset*: As sequências de nucleotídeo foram recuperadas do GenBank com um *script* escrito em linguagem Python através das palavras-chave "Leptospira [organism]". As sequências de cada isolado foram indexadas em um banco de dados com base na anotação quanto à sua espécie. As sequencias dos loci do MLST (7 loci) e dos 3 conjuntos do PubMLST (21 loci) para *Leptospira* sp. foram recuperadas de seus respectivos repositórios. As sequências dos genes *rrs*, *rpoB*, *SecY*, *LigB*, e *LipL41* foram extraídas do genoma da *Leptospira interrogans* sorovar copenhageni L1-130. Para cada um dos 33 loci, buscas através do BLASTn foram realizadas para a identificação das suas sequências homólogas em cada isolado indexado no nosso banco de dados. Para cada *loci* foi realizado o alinhamento múltiplo das sequências referentes a cada isolado, sendo o arquivo indexado em uma tabela, e suas posições não conservadas, filtradas.

Aplicação dos algoritmos de mineração: foram utilizados 4 diferentes algoritmos de classificação: J48 (baseado em árvore de decisão), *Multilayer Perceptron* (baseado em redes neurais), Naive Bayes (RISH, 2001) e SVM (COLLOBERT; BENGIO, 2004), usando as implementações presentes no programa Weka (*Waikato Environment for Knowledge Analysis*), sendo também feita a comparação da eficiência destes algoritmos quando utilizados no *Adaboost*.

Avaliação dos modelos: para a avaliação dos métodos de classificação foi considerada a capacidade de cada abordagem de identificar diferentes espécies. Os modelos de classificação gerados pelo Weka foram avaliados através de validação cruzada e *percent split* de 80%.

3. RESULTADOS E DISCUSSÃO

Atualmente são conhecidas 21 espécies de *Leptospira* e, através de *script*, 19 delas foram recuperadas do GenBank e registradas no nosso banco de dados, sendo elas: *L. interrogans, L. kirschneri, L. santarosai, L. borgpetersenii, L. noguchii, L. weilli, L. alstonii, L. alexanderi* (patogênicas)*, L. kmetyi, L. broomii, L. licerasiae, L. wolffii, L. fainei, L. inadai* (intermediádias)*, L. meyeri, L. terpstrae, L. wolbachii, L. yanagawae, L. vanthielli* (saprófitas), totalizando 600 genomas. As sequências gênicas também foram recuperadas dos repositórios, totalizando 33 loci.

Através do BLASTn, foi realizado um alinhamento múltiplo entre as sequências de cada isolado e de cada um dos 33 loci. As posições polimórficas dos nucleotídeos foram filtradas e indexadas em uma tabela para serem avaliadas por algoritmos de mineração de dados presentes no *software* Weka. Dos quatro algoritmos propostos para a classificação de *Leptospira*, J48, SVM e Naive Bayes

foram executados com sucesso para a análise de classificação. O algoritmo *Multilayer Perceptron* não foi eficiente em classificar, provavelmente pela grande quantidade de parâmetros inseridos. A figura 1 apresenta a acurácia na classificação de 600 espécies de *Leptospira* para cada um dos algoritmos computacionais utilizados.

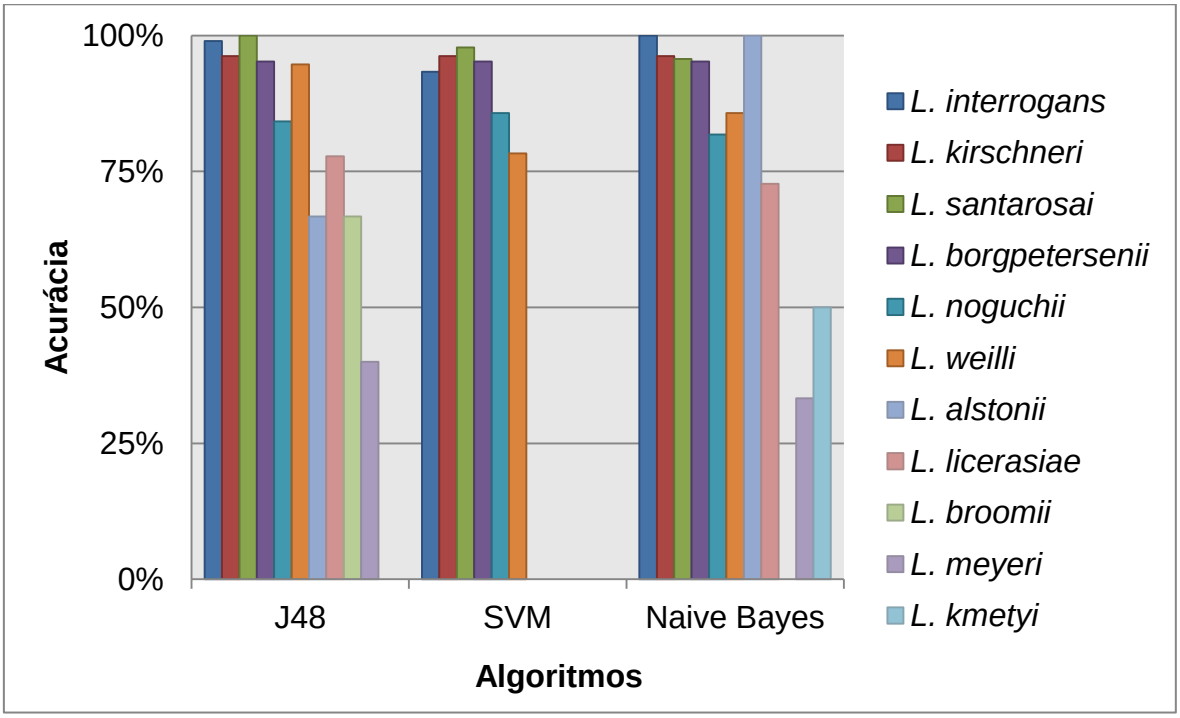

Figura 1. Acurácia de três algoritmos computacionais - J48, SVM e Naive Bayes - em classificar 600 isolados de *Leptospira* com base em 33 loci.

A classificação por algoritmos computacionais levou em conta a quantidade de sequências de *Leptospira* disponíveis no banco de dados, de modo que, espécies que possuíam poucas sequências disponíveis para comparação (*L. kmetyi, L. inadai*, *L. fainei, L. terpstrae, L. wolbachii* e *L. vanthielli*, as quais somam apenas 16 genomas*)*, foram removidas das análises. A figura 1 nos mostra que J48 e Naive Bayes classificaram 10 espécies e SVM classificou apenas 6 espécies. A média da porcentagem de acurácia de cada algoritmo foi 74,6%, 49,7% e 73,7% para J48, SVM e Naive Bayes respectivamente (dados não mostrados). Comparando esses resultados podemos perceber que J48 se mostrou mais apropriado para classificar diferentes capas de *Leptospira*, pois, além de classificar um maior número de espécies (comparado a SVM), classificou também com maior acurácia cada uma delas. Estes algoritmos também serão utilizados para classificar isolados de *Leptospira* de diferentes níveis hierárquicos como sorogrupo, sorovar e cepa. Para a comparação dos atuais modelos de classificação molecular, buscas por BLAST também serão realizadas para cada loci, sendo os resultados dos loci de MLST comparados com as tabelas de STs de seus respectivos repositórios, e os genes comparados com os melhores *hits* no Genbank.

4. CONCLUSÕES

Foi possível observar que, baseado na mineração de dados de 600 genomas e 33 *loci* de genes de *Leptospira*, a aplicação do algoritmo computacional J48 foi o mais apropriado na classificação de diferentes espécies

de *Leptospira*, pois, além de classificar um maior número de espécies, classificou também com maior acurácia cada uma delas. Desta forma, J48 pode se utilizado para auxiliar na classificação de isolados de *Leptospira*.

5. REFERÊNCIAS BIBLIOGRÁFICAS

ADLER, B.; MOCTEZUMA, A. D. L. P. Leptospira and leptospirosis. **Veterinary Microbiology**, v. 140, p. 287–296, 2010.

AHMED, N. et al. Multilocus sequence typing method for identification and genotypic classification of pathogenic Leptospira species. **Annals of clinical microbiology and antimicrobials**, v. 5, p. 28, jan. 2006.

CERQUEIRA, G. M. et al. Bioinformatics describes novel Loci for high resolution discrimination of leptospira isolates. **PloS one**, v. 5, n. 10, p. e15335, jan. 2010.

COLLOBERT, R.; BENGIO, S. **Links between Perceptrons , MLPs and SVMs**Proceedings of the 21 st International Conference on Machine Learning. **Anais**...Banff, Canada: 2004

MAIDEN, M. C. et al. 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**, v. 95, n. March, p. 3140–3145, 1998.

MOREY, R. E. et al. Species-specific identification of Leptospiraceae by 16S rRNA gene sequencing. **Journal of Clinical Microbiology**, v. 44, n. 10, p. 3510–3516, 2006.

RISH, I. IBM Research Report An empirical study of the naive Bayes classifier. **Computer Science**, v. 22230, 2001.

ROMERO, E. C.; YASUDA, P. H. 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**, v. 101, n. June, p. 373–378, 2006.

SCOLA, B. LA et al. Partial rpoB gene sequencing for identification of Leptospira species. **FEMS Microbiology Letters**, v. 263, n. Table 1, p. 142–147, 2006.

THAIPADUNGPANIT, J. et al. A dominant clone of Leptospira interrogans associated with an outbreak of human leptospirosis in Thailand. **PLoS Neglected Tropical Diseases**, v. 1, n. 1, p. 1–6, 2007.

WHO. Leptospirosis: an emerging public health problem. **The Weekly Epidemiological Record (WER)**, p. 45–52, 2011.