ARSENAL: Antimicrobial ReSistance prEdictioN by a mAchine Learning method

Ulysse Guyet[1,2], Léa Bientz[3], Véronique Dubois[3], Jie Feng[4], Jacques Corbeil[5,6], Alexis Groppi[1,2], and Macha Nikolski[1,2]

[1] Univ. Bordeaux, CNRS, IBGC, UMR 5095, Bordeaux, 33077, France
[2] Univ. Bordeaux, Centre de Bioinformatique de Bordeaux (CBiB), Bordeaux, 33076, France
[3] MFP, CNRS 5234, Université de Bordeaux, Bordeaux, F-33076, France
[4] State Key Laboratory of Microbial Resources, Institute of Microbiology, Chinese Academy of Sciences, Beijing, 100101, China
[5] Research Center in Infectious Diseases, CHU de Québec-Laval University Research Center and Department of Molecular Medicine and Big Data Research Centre, Faculty of Medicine, Laval University, Quebec City, QC, Canada
[6] Department of Molecular Medicine, Laval University, Quebec City, QC, Canada

Abstract. Antimicrobial resistance (AMR) has become a major public health concern due to the rapid emergence of multidrug-resistant bacteria, causing serious problems for the prevention and treatment of persistent infections. Development of algorithms for phenotypic variation prediction, such as AMR, could be of major clinical importance, more reliable and efficient compared to traditional phenotyping, and could contribute to the discovery of previously unknown AMR pathways. Significant increase of the available sequencing and associated phenotypic data in recent years creates the basis for the development of such methods. Here, we developed a machine learning method -ARSENAL- for predicting the minimum inhibitory concentration (MIC) of several antibiotics based on genomic data. ARSENAL relies on one hand on the sequence (k-mers), and on the other hand on the genome structure (gene composition) and the gene orthology links between the strains of the same species. Functional interpretation of the most predictive features confirmed the biological relevance of the ARSENAL model.

Keywords: Machine learning · Anti-microbial resistance · Genomic data · *Pseudomonas aeruginosa* · *Streptococcus pneumoniae*.

This research has received funding from the JPIAMR under the grant agreement ANR-19-JAMR-0006.

1 Introduction

Antimicrobial resistance (AMR) is a growing health threat responsible for an estimated 700 000 deaths per year and is expected to cause 10 million deaths per year by 2050 [17]. Appropriate antibiotic therapy improves patient healing outcomes and is a key factor in preventing the emergence of antibiotic resistance [12, 23].

Antibiotic susceptibility testing (AST) from bacterial culture is the current clinical practice for assessing drug resistance by the determination of the minimum inhibitory concentration (MIC) corresponding to the lowest concentration of a specific antibiotic that inhibits bacterial growth. This method, fastidious and long (between 24h and 72h), presents a certain number of error sources, in particular during the preparation of the inoculum or due to the culture conditions [2]. Furthermore, AST is only applicable to cultivable bacteria, which excludes analysis of the emergence and spread of antimicrobial resistance in diverse and complex microbial communities with large fractions of currently uncultured bacteria [4].

Constant improvements in sequencing techniques and computational methods had allowed the rapid and efficient characterization of genomes and their genes. Availability of such large annotated genomic datasets combined with MIC data enables the development of computational methods to predict the level of resistance. Existing methods for resistance prediction mainly fall in the machine learning category that are based on k-mer statististics gathered from full genomes. These methods allow either to explicitly highlight associations between genotype and binary phenotype (resistant/susceptible), or to establish a more or less precise prediction of the MIC value depending on the organism and antibiotics [7, 15, 24, 10]. Nevertheless, neither of these two types of methods allows linking small-scale genomic information (kmer) with larger-scale genomic sequence (gene) carrying functional information. To fill this gap and enable more straightforward biological interpretations, we have developed a machine learning method (ARSENAL) based on genomic data to predict MICs and able to make a direct link between phenotype and new potential biomarkers of resistance.

2 Methods

2.1 Genome assembly and annotation

In this study, 1312 genomes of *Streptococcus pneumoniae* strains were newly sequenced as 150 bp paired-end reads. These genomes were assembled using the SPAdes genome assembler v.3.13.0 [1] with the following parameters '–careful -t 8 -m 48 -k 21,33,45,55,63,77 –cov-cutoff auto'. Assembly quality and genome completeness metrics were computed using QUAST v4.6.3 [8] and BUSCO v5.1.2 [22]. Automatic gene prediction as well as structural and functional annotation of the genomes was performed using PATRIC database API [26]. Specifically, each gene has been assign to a PLFam (PATRIC Local protein family), a group of genus-specific genes which share the same function and high sequence homology.

2.2 Minimum Inhibitory Concentration assay

MICs were determined according EUCAST recommendations by broth microdilution, *i.e.* 2-fold by 2-fold dilutions of the antibiotic, and a standardized bacterial inoculum of 5 x10^5 CFU/mL in a Mueler-Hinton liquid medium were added. Cultures were then incubated at 37°C during 20h±4h and MIC were defined as the concentration of antibiotic that inhibits any visible bacterial growth (*i.e.* the first non cloudy well).

2.3 Machine learning prediction pipeline

For each genome annotated with PATRIC, each PLFam was divided into a set of nonredundant overlapping nucleotide 6-mers using the k-mer counting program KMC [5] (Fig. 1). For each PLFam, a matrix was built in which the k-mers and MICs are treated as features for each genome. Each row in the specific PLFam matrix contains the k-mers counts for a genome, as well as the MIC for a single antibiotic. Then, for each PLFam, a MIC prediction model was built using a XGBoost [3] regressor predicting linearized MICs. Among the advantages of this method, we can highlight its scalability and its built-in ability to perform feature selection. The predictions of each XGBoost model of each PLFam are then gathered in a single table containing in rows the genomes and in columns the predictions of MIC for a given PLFam. A random forest model was then trained, taking as input this table of prediction data and giving as output a prediction of MIC value per genome. The predictions of each model of a PLFam are then gathered in a same table containing in row the genomes and in columns the predictions of MIC for a given PLFam. A random forest model was then trained, taking as input this table of prediction data and giving as output a prediction of MIC value per genome. Models were trained and evaluated in a 5x genome-distance-based cross-validation scheme, such that genome distance was maximized between the test sets of folds (as described in [13]).

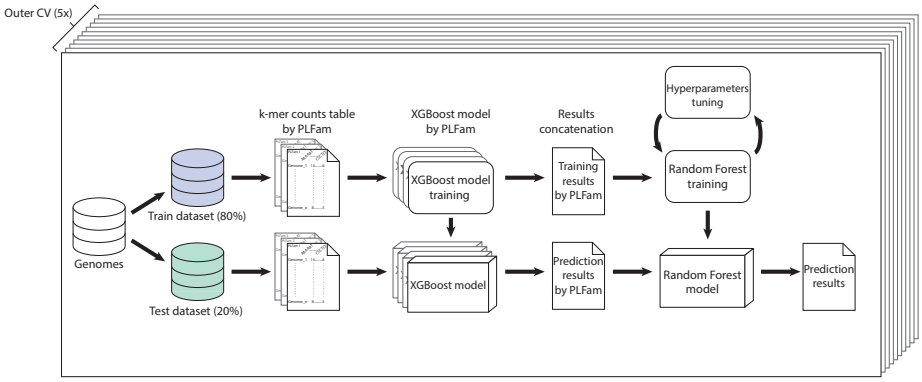

Fig. 1. ARSENAL pipeline description.

2.4 Determination of feature importance

Shapley additive values [21], additive explanations assigning a relative importance to each feature, were calculated and provided a framework to better understand the contribution of each PLFam to the MIC prediction.

3 Preliminary results

3.1 Genome assemby and annotation

Assembled *Streptococcus pneumoniae* genomes are highly contiguous and complete, with an average contig N50 of 101 Mb, an average L50 of 7.6 and an average completeness of 99.97%. Automatic gene prediction with PATRIC API resulted in functional annotation for 83% of predicted genes in average.

3.2 Prediction accuracy

In order to evaluate the performance of our prediction pipeline, we considered the accurracy, evaluated for each β-lactam antibiotic tested on the 1312 genomes of *Streptococcus pneumoniae* (see Table 1). Cefuroxime had the highest accuracy (88%), and cefepime had the lowest accuracy (74.9%) (Table 1).

Table 1. Accuracies of the MIC prediction pipeline for β-lactam antibiotics based on 1312 *Streptococcus pneumoniae* genomes. The table depicts the accuracy within ± 1 2-fold dilution step of the laboratory-derived MIC.

Antibiotic Class	Penicillins		Cephalosporins			Carbapenem
Antibiotic	Penicillin	Amoxicillin	Cefuroxime	Cefepime	Ceftriaxone	Imipinem
Accuracy	0.76	0.807	0.88	0.749	0.795	0.878

3.3 Feature importance

To estimate the extent to which each feature (in our case, PLFam gene families) contributed to the prediction of the model, we used Shapley Additive Explanations to calculate the local feature importance for each observation and ranked the genes by importance for the prediction model. In the case of AMR prediction of *Streptococcus pneumoniae* strains to penicillin, we obtained 336 genes with a Shapley value higher than 0.10. Among the genes of interest, we identified the genes encoding the 3 penicillin-binding proteins of *Streptococcus pneumoniae*, ranked 10th, 39th, and 46th. These proteins are the target enzymes of β-lactam antibiotics and have been previously described as highly altered in clinical isolates with a mosaic genetic structure indicating interspecies gene transfer followed by recombination events [6, 11]. We also found a significant number of mobile genetic elements (9 among the 200 most important genes). Although it is necessary to identify precisely the nature of these sequences, mobile genetic elements are described as largely involved in the capture, accumulation and dissemination of resistance genes [18, 9]. We also identified, in 13th position, *metG* which encodes a methionine-tRNA ligase and whose specific mutations reduce the efficiency of some antibiotics [27]. The *lytA* gene, in the 12th position, encodes the autolysin LytA, an autolytic enzyme inducing bacterial cell

autolysis when activated by cell wall acting antibiotics [25]. However, mutation of this gene can prevent autolysis and make bacteria tolerant to antibiotics that inhibit cell wall synthesis [20, 19, 16, 14].

4 Conclusion and future work

In the literature, most studies that aim to predict antibiotic resistance from genomic data only provide a binary prediction (resistant or susceptible). Here, we presented a new method that can predict a precise level of resistance (MIC value) in bacterial strains from their genome but also allows determining the genes that have the most impact on the model prediction. Although we have been able to identify a number of these genes whose functions are related to antimicrobial resistance, the function of many of the output genes of the model remains to be characterized (e.g. by site-directed mutagenesis) and may allow us to identify new biomarkers of resistance. Our method has been applied to strains of *Streptococcus pneumoniae* (gram-negative) in the case of resistance to beta-lactams, and we plan to test it in the case of other families of antibiotics as well as on strains of *Pseumonas aeruginosa* (gram-positive).

References

1. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al.: Spades: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology **19**(5), 455–477 (2012)
2. Benkova, M., Soukup, O., Marek, J.: Antimicrobial susceptibility testing: currently used methods and devices and the near future in clinical practice. Journal of Applied Microbiology **129**(4), 806–822 (2020)
3. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp. 785–794 (2016)
4. D'Costa, V.M., McGrann, K.M., Hughes, D.W., Wright, G.D.: Sampling the antibiotic resistome. Science **311**(5759), 374–377 (2006)
5. Deorowicz, S., Kokot, M., Grabowski, S., Debudaj-Grabysz, A.: Kmc 2: fast and resource-frugal k-mer counting. Bioinformatics **31**(10), 1569–1576 (2015)
6. Dowson, C., Hutchison, A., Spratt, B.: Extensive re-modelling of the transpeptidase domain of penicillin-binding protein 2b of a penicillin-resistant south african isolate of streptococcus pneumoniae. Molecular microbiology **3**(1), 95–102 (1989)
7. Drouin, A., Letarte, G., Raymond, F., Marchand, M., Corbeil, J., Laviolette, F.: Interpretable genotype-to-phenotype classifiers with performance guarantees. Scientific reports **9**(1), 1–13 (2019)
8. Gurevich, A., Saveliev, V., Vyahhi, N., Tesler, G.: Quast: quality assessment tool for genome assemblies. Bioinformatics **29**(8), 1072–1075 (2013)
9. Harbottle, H., Thakur, S., Zhao, S., White, D.: Genetics of antimicrobial resistance. Animal biotechnology **17**(2), 111–124 (2006)
10. Kim, J., Greenberg, D.E., Pifer, R., Jiang, S., Xiao, G., Shelburne, S.A., Koh, A., Xie, Y., Zhan, X.: Vampr: Va riant m apping and p rediction of antibiotic r esistance via explainable features and machine learning. PLoS computational biology **16**(1), e1007511 (2020)

11. Laible, G., Spratt, B., Hakenbeck, R.: Interspecies recombinational events during the evolution of altered pbp 2x genes in penicillin-resistant clinical isolates of streptococcus pneumoniae. Molecular microbiology **5**(8), 1993–2002 (1991)
12. Lee, C.R., Cho, I.H., Jeong, B.C., Lee, S.H.: Strategies to minimize antibiotic resistance. International journal of environmental research and public health **10**(9), 4274–4305 (2013)
13. Lv, J., Deng, S., Zhang, L.: A review of artificial intelligence applications for antimicrobial resistance. Biosafety and Health **3**(01), 22–31 (2021)
14. Mitchell, L.S., Tuomanen, E.I.: Molecular analysis of antibiotic tolerance in pneumococci. International journal of medical microbiology **292**(2), 75–79 (2002)
15. Nguyen, M., Long, S.W., McDermott, P.F., Olsen, R.J., Olson, R., Stevens, R.L., Tyson, G.H., Zhao, S., Davis, J.J.: Using machine learning to predict antimicrobial mics and associated genomic features for nontyphoidal salmonella. Journal of clinical microbiology **57**(2), e01260–18 (2019)
16. Normark, B.H., Normark, S.: Antibiotic tolerance in pneumococci. Clinical Microbiology and Infection **8**(10), 613–622 (2002)
17. O'Neill, J., et al.: Review on antimicrobial resistance. Antimicrobial resistance: tackling a crisis for the health and wealth of nations **2014**(4) (2014)
18. Partridge, S.R., Kwong, S.M., Firth, N., Jensen, S.O.: Mobile genetic elements associated with antimicrobial resistance. Clinical microbiology reviews **31**(4), e00088–17 (2018)
19. Ronda, C., García, J.L., García, E., Sánchez-Puelles, J.M., López, R.: Biological role of the pneumococcal amidase: cloning of the lyta gene in streptococcus pneumoniae. European journal of biochemistry **164**(3), 621–624 (1987)
20. Sánchez-Puelles, J.M., Ronda, C., Garcia, J.L., Garcia, P., Lopez, R., Garcia, E.: Searching for autolysin functions: characterization of a pneumococcal mutant deleted in the lyta gene. European journal of biochemistry **158**(2), 289–293 (1986)
21. Shapley, L.S., Kuhn, H., Tucker, A.: Contributions to the theory of games. Annals of Mathematics studies **28**(2), 307–317 (1953)
22. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M.: Busco: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics **31**(19), 3210–3212 (2015)
23. Stoesser, N., Batty, E., Eyre, D., Morgan, M., Wyllie, D., Del Ojo Elias, C., Johnson, J., Walker, A., Peto, T., Crook, D.: Predicting antimicrobial susceptibilities for escherichia coli and klebsiella pneumoniae isolates using whole genomic sequence data. Journal of Antimicrobial Chemotherapy **68**(10), 2234–2244 (2013)
24. Tan, R., Yu, A., Liu, Z., Liu, Z., Jiang, R., Wang, X., Liu, J., Gao, J., Wang, X.: Prediction of minimal inhibitory concentration of meropenem against klebsiella pneumoniae using metagenomic data. Frontiers in Microbiology **12** (2021)
25. Tomasz, A., Albino, A., Zanati, E.: Multiple antibiotic resistance in a bacterium with suppressed autolytic system. Nature **227**(5254), 138–140 (1970)
26. Wattam, A.R., Abraham, D., Dalay, O., Disz, T.L., Driscoll, T., Gabbard, J.L., Gillespie, J.J., Gough, R., Hix, D., Kenyon, R., et al.: Patric, the bacterial bioinformatics database and analysis resource. Nucleic acids research **42**(D1), D581–D591 (2014)
27. Yi, H., Lee, H., Cho, K.H., Kim, H.S.: Mutations in metg (methionyl-trna synthetase) and trmd [trna (guanine-n1)-methyltransferase] conferring meropenem tolerance in burkholderia thailandensis. Journal of Antimicrobial Chemotherapy **73**(2), 332–338 (2018)