The 8th IEEE International Conference on E-Health and Bioengineering - EHB 2020
Grigore T. Popa University of Medicine and Pharmacy, Web Conference, Romania, October 29-30, 2020

Using Additive Bayesian Networks and Association Rules in Antimicrobial Resistance Analysis

Raluca Mureșan[1], Claudia Zaharia[2], Daniela Zaharie[1]
[1]Department of Computer Science, Faculty of Mathematics and Computer Science, West University of Timișoara
Timișoara, Romania, raluca.muresan@e-uvt.ro, daniela.zaharie@e-uvt.ro
[2]Department of Mathematics, Faculty of Mathematics and Computer Science, West University of Timișoara
Timișoara, Romania, claudia.zaharia@e-uvt.ro

***Abstract*— The aim of this paper is to analyze the effectiveness of combining two data-driven approaches, additive Bayesian networks (ABN) and association rules mining (ARM), in identifying relevant patterns of antimicrobial resistance (AMR). The main idea is to use information provided by ARM as prior knowledge in the inference of ABNs describing relationships between antimicrobials involved in resistance patterns. The results obtained for a dataset containing Escherichia coli isolates illustrate that by combining the two approaches one can better explain AMR patterns of association than by using only one of the methods.**

Keywords—antimicrobial resistance; additive bayesian networks; association rules; Escherichia coli

I. INTRODUCTION

Antimicrobial resistance (AMR) is one of the biggest global public health challenges of our time, with millions of people getting drug-resistant infections each year. AMR thus places a tremendous burden on healthcare systems and society. Gaps in our understanding of the complexity of AMR impede efficient treatment and make prevention very difficult.

The presence of multiple resistance determinants within bacterial isolates can occur due to various biological and evolutionary mechanisms, often resulting in multidrug resistance (MDR) [5]. It is therefore of interest to assess systematic co-dependencies between resistances to find potential factors of intervention. Exploration of MDR patterns is extensively studied in the literature using different statistical models, e.g. loglinear models, Bayesian networks, Markov networks, dynamical Bayesian networks, generalized linear models (see [4] and the references therein). Another technique that was recently proposed for the investigation of MDR patterns is association rule mining [4]. Identification of relevant patterns of association in the context of multidrug resistance is challenging as there are many ways of ranking the association rules extracted from data. A methodological approach for choosing interestingness measures to be used in ranking associations between antimicrobial drugs which induce resistance is presented in [21].

In recent years, Additive Bayesian Networks (ABN) have emerged as a promising technique for uncovering complex interactions between the variables in a data set, and in particular for assessing patterns of association between drug resistances ([9, 13, 17]). ABN modeling is a new type of graphical approach for the analysis of complex systems that is increasingly applied in areas like genetics, systems biology, livestock production and epidemiology [6, 14].

It is worth pointing out that when there are many potential interactions to be inferred from a limited volume of sparse data, as is usually the case when analyzing resistance patterns, it can pose difficulties for the learning algorithm. In this case, many authors advocate for the inclusion of complementary information in the learning process, in the form of domain-specific knowledge (see [8, 10, 20]). Djebbari and Quackenbush [8] showed that using prior knowledge as network seeds greatly improves the ability of Bayesian Network analysis to learn interaction networks and argued that such seeds should combine diverse sources of data and information.

In this paper, we propose an approach that integrates the findings of association rule mining as prior knowledge into the ABN model, to show that combining these two techniques can lead to a better understanding of MDR patterns. We build upon the framework of our previous paper [21] and re-examine the data and results therein in light of this new approach.

II. DATABASE AND METHODS

A. Database

The data was extracted from the National Antimicrobial Resistance Monitoring System for Enteric Bacteria (NARMS) database [18] and consists of 329 E. coli isolates obtained from humans between 1996-2016. Each isolate was classified as either resistant or susceptible based on published minimum inhibitory concentration breakpoints, to the following 10 antibiotics: Ampicillin (AMP), Amoxicillin-Clavulanic Acid (AUG), Ceftriaxone (AXO), Chloramphenicol (CHL), Ciprofloxacin (CIP), Nalidixic Acid (NAL), Gentamicin (GEN), Streptomycin (STR), Trimethoprim-Sulfamethoxazole (COT) and Tetracycline (TET). All isolates were resistant to at least one antibiotic.

B. Association rules

Association rule mining [1, 2] is an unsupervised machine learning technique used to identify relationships between items in a transaction database. Given a set of items, *I*, and a set of

IEEE 978-1-7281-8803-4/20/$31.00 ©2020 IEEE

transactions, D, where each transaction is a subset of I, an association rule is an implication of the form $X \rightarrow Y$, where X and Y are disjoint sets of items from I. In the case study presented in this paper, I is the set of antimicrobials, and each of the 329 isolates with the subset of antimicrobials to which it was classified as resistant represents a transaction. Three measures are commonly used to assess the quality of a rule. The support of the rule $X \rightarrow Y$ is the proportion of transactions from D that contain both X and Y. The confidence of the rule is defined as the conditional probability of Y, given X. The lift is defined as $P(X,Y)/(P(X) \cdot P(Y))$ and quantifies how many times more often X and Y occur together than would be expected under the assumption of statistical independence. The association rule mining process has two steps: first, all itemsets with support greater than or equal to a given threshold are found (these are called frequent itemsets); then, the frequent itemsets are used to generate rules.

Association rules proved to be a valuable tool in AMR analyses [4, 21]. The technique is well suited for handling sparse data and does not require any distributional assumptions. There are, however, limitations which must be taken into account. As rule mining takes place in a support – confidence framework, the extraction of rare rules can be problematic [16]. Namely, suppose that strains resistant to antimicrobial A are found very rarely, but when they do occur, they are resistant to antimicrobial B as well. Such an association may be worth investigating from a practical standpoint, but by using support-based filtration to select the candidate ruleset, such rare rules can be easily missed. If, however, one intends to find rare but interesting rules by setting a low minimum support threshold, this will also create many unwanted rules, which are in fact just spurious item associations.

C. Additive Bayesian Networks

ABN are a form of graphical statistical modelling which seeks to infer an appropriate probabilistic model which best describes the joint probability structure - co-dependencies - within observed data, distinguishing between direct and indirect associations [15]. It is especially suited for analyses of complex data where the variables are highly correlated. ABN are an extension of generalized linear regression models (GLM) to the multivariate case where all relations between the variables are taken into consideration [13]. Several advantages come from this: in comparison to classical regression models no dimension reduction techniques are used making the results more transparent and easy to interpret; the landscape of considered models is much larger in the case of ABN enabling the discovery of more complex relations; the Yule-Simpson paradox, which states that once more variables are added to the model an apparent relationship between variables can disappear or even be reversed, is avoided [17]. Moreover, this graphical procedure provides results which are easy to visualize and interpret, making the relations between the variables straightforward to understand. One disadvantage of ABN is the computational burden this technique involves since it searches for an optimal structure among all directed acyclic graphs (DAGs).

An ABN consists of a DAG and a set of parameters. Therefore, the method consists of two interconnected parts: (1) *learning the structure* and (2) *estimating the parameters*. The graph consists of a set of nodes and directed edges or arcs. Node i is said to be a parent of node j if there is an edge from i to j; node j is said to be the child of node i. The nodes of the graph represent the variables and for each node a GLM is considered where the covariates are the parents. In the analysis conducted in this paper the GLMs are logistic regressions since all the variables are dichotomous.

The structure learning part assumes an exact or a heuristic search among the landscape of all possible DAGs and chooses the one that best fits the data according to an information theoretic metric. For the present data set an exact search was performed, using the abn package in R [12], the chosen scoring metric being the Bayesian Information Criterion (BIC). To avoid potential overfitting due to data separation or sparsity, model averaging was performed using Markov Chain Monte Carlo (MCMC) generating samples from the posterior distribution of the most supported DAGs. More precisely, four chains were generated starting from the globally optimal DAG, each with a length of 50000, a burn-in phase of 5000 steps and a thinning step of 100 to avoid autocorrelation, using the mcmcabn package in R [11]. Convergence was checked by visual inspection of trace plots and using the Gelman-Rubin convergence diagnostic [7]. Averaging among the four chains, a majority consensus DAG was built by removing any arc with a support less than 50%.

D. Using association rules information in ABN modelling

ABN modelling is a data driven approach, but can include expert knowledge in various ways, thus becoming a semi-supervised method. In order to integrate prior knowledge into the ABN model, we use the approach of Werhli and Husmeier [20]. The prior information is encoded by means of a matrix B with entries $B_{ij} \in [0,1]$, where B_{ij} represents the degree of belief that there is a directed arc from node i to node j, the belief being stronger as B_{ij} gets closer to 1. The agreement between a given network and the available prior knowledge is quantified using an energy function. A prior distribution over network structures is defined, based on the energy and a hyperparameter $\beta \in (0,\infty)$ that controls the strength of the influence of the prior knowledge relative to the data, larger values indicating a greater influence. This distribution is then used in the MCMC sampling scheme.

Here, we used the information regarding resistance patterns extracted from the data set using association rule mining to guide the search for the ABN model. Specifically, in the prior knowledge matrix, B_{ij} was set to be the confidence of the rule $i \rightarrow j$ (all rules containing exactly one antecedent and one consequent are used). Model averaging was performed using mcmcabn as above, with a user defined prior given by the matrix B and the hyperparameter β set to 1.

III. RESULTS

Since the results should be interpreted in the context of the prevalence of resistance to each antibiotic, a preliminary

analysis of the dataset has been conducted leading to the prevalence values presented in Table I.

TABLE I. PREVALENCE OF RESISTANCE TO THE 10 ANTIBIOTICS IN THE DATASET

	AMP	AUG	AXO	CHL	CIP	NAL	GEN	STR	COT	TET
No of strains	106	14	13	60	10	60	15	147	48	200
%	32.22	4.26	3.95	18.24	3.04	18.24	4.56	44.68	14.59	60.79

The globally optimal ABN model of susceptibility patterns found using the exact search method, supposed to best fit the data, has 16 arcs (data not shown). Using this DAG as the starting point for the MCMC model averaging and eliminating the arcs with support less than 50% lead to the model presented in Fig. 1. Only 8 of the arcs were retrieved in the majority consensus DAG, indicating the presence of overfitting in the initial model. Table II shows the odds ratio of each arc presented as "parent→child", the 95% Wald-type confidence intervals and arc support (percentage of generated structures containing the arc) for the model given in Fig. 1. It also contains the values of support, confidence and lift for the corresponding association rules.

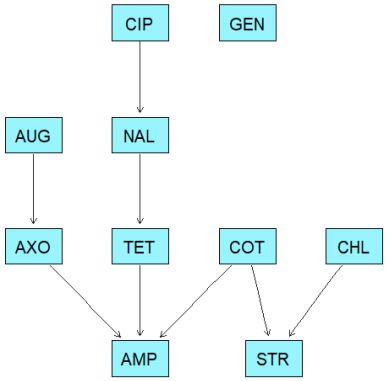

Fig. 1. ABN model after adjustement for overfitting using MCMC with the globally optimal DAG as starting point.

TABLE II. ODDS RATIO, 95% CONFIDENCE INTERVALS, ARC SUPPORT (%), RULE SUPPORT, CONFIDENCE AND LIFT FOR THE MODEL IN FIG. 1

Arc (p→c)	OR (95% CI)	Arc Support (%)	Rule support	Rule confidence	Rule lift
COT→AMP	9.03 (4.37, 18.64)	74.75	0.09	0.65	2.00
AXO→AMP	561.16 (0.48, 6.51e+05)	74.25	0.04	1.00	3.10
TET→AMP	0.29 (0.17, 0.52)	55.24	0.15	0.25	0.78
CIP→NAL	47.47 (5.83, 386.53)	54.93	0.03	0.90	4.94
NAL→TET	0.23 (0.12, 0.41)	54.55	0.06	0.32	0.52
COT→STR	4.66 (2.17, 10.02)	52.87	0.12	0.79	1.77
CHL→STR	4.44 (2.28, 8.64)	52.56	0.14	0.77	1.72
AUG→AXO	259.82 (51.07, 1321.87)	51.12	0.03	0.71	18.08

Strong positive associations were found between resistances to COT and AMP, CIP and NAL, COT and STR, CHL and STR, and AUG and AXO. Resistances to AXO and AMP also appear to be very strongly positively associated, with an odds ratio of 561, although the 95% confidence interval for the odds ratio is very wide and includes the value 1, which would deem the association not statistically significant. An explanation for this pattern comes when looking further at the interestingness measures of the rule AXO→AMP. The confidence of the rule is 1, meaning that all isolates resistant to AXO in the dataset were also resistant to AMP, and the lift is 3.1 – consequently, the association is indeed a relevant one, while the wide confidence interval is a consequence of numerical instability due to data separation.

In the paper [21], a minimum support threshold of 0.06 was imposed to prune the ruleset in order to ensure a false discovery rate below 5%. As pointed out above, this may filter out valuable rare rules along with spurious associations. This appears to be the case here, the associations between resistances to AXO and AMP, CIP and NAL, AUG and AXO are very strong but occur very seldom, as individual resistance to any of AXO, AUG and CIP is, in itself, still rare (see Table I). Thus, they were missed when using only association rules to analyze the dataset but discovered when using association rules alongside ABN.

Additionally, there appear to be negative associations between resistances to TET and AMP, and NAL and TET. The corresponding rules have low confidence and lift less than 1. When using the approach described in Section II.D, incorporating the knowledge extracted from the association rules as arc priors, MCMC model averaging and elimination of the arcs occurring in less than 50% of the generated DAGs lead to the model presented in Fig. 2 and Table III.

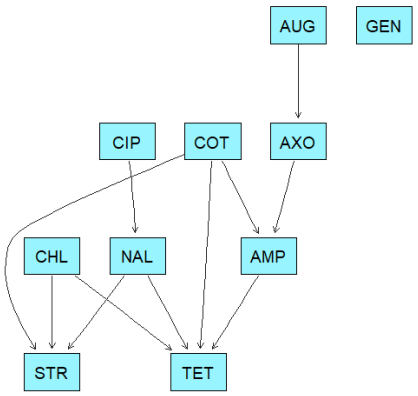

Fig. 2. ABN model obtained by performing MCMC with an empty DAG as a starting point and priors for each arc given by the corresponding association rule confidence.

TABLE III. ODDS RATIO, 95% CONFIDENCE INTERVALS, ARC SUPPORT (%), RULE SUPPORT, CONFIDENCE AND LIFT FOR THE MODEL IN FIG. 2

Arc (p→c)	OR (95%CI)	Arc Support (%)	Rule support	Rule confidence	Rule lift
COT→TET	8.17 (3.19, 20.92)	79.70	0.12	0.81	1.33
AMP→TET	0.18 (0.10, 0.32)	79.17	0.15	0.47	0.78
NAL→TET	0.13 (0.07, 0.26)	77.82	0.06	0.32	0.52
AXO→AMP	720.54 (0.22, 2.41e+06)	74.35	0.04	1.00	3.10
COT→STR	5.47 (2.50, 11.99)	70.72	0.12	0.79	1.77

CHL→STR	4.1 (2.10, 7.98)	65.67	0.14	0.77	1.72
COT→AMP	5.93 (3.11, 11.32)	65.43	0.09	0.65	2.00
CIP→NAL	47.47 (5.83, 386.53)	61.27	0.03	0.90	4.94
NAL→STR	0.31 (0.15, 0.64)	58.27	0.05	0.27	0.60
AUG→AXO	259.82 (51.07, 1321.87)	52.00	0.03	0.71	18.08
CHL→TET	3.46 (1.55, 7.72)	51.00	0.15	0.82	1.34

Beside the previously discovered patterns, this approach highlights significant positive associations between resistances to COT and TET, CHL and TET, and a significant negative association between resistances to NAL and STR. Among these additional patterns, the first two were also found in [21] and have been documented clinically ([3], [19]).

IV. Conclusions

The analysis conducted in this paper shows the advantages of combining association rule mining and additive Bayesian networks for the discovery of antimicrobial resistance patterns. Association rules can be used both as prior knowledge to guide the search for the ABN model and as means to aid the interpretation of the ABN results. By using the two methods together, one can obtain information on both the strength and frequency of associations, overcoming each method's specific limitations. As seen above, using the confidence of association rules as prior knowledge in the ABN model lead to the discovery of supplementary patterns with clinical relevance. Further work aims to provide a formal validation of this approach through a theoretical study.

Acknowledgment

The research was carried out in the frame of the project Bioeconomic approach to antimicrobial agents - use and resistance financed by UEFISCDI by contract no. 7PCCDI/ 2018, cod PN-III-P1-1.2-PCCDI-2017-0361.s

References

[1] R. Agrawal, T. Imielinski, A. Swami, Mining association rules between sets of items in large databases, in: Proceedings of 1993 ACM-SIGMOD International Conference on Management of Data, Washington, DC, May 1993, pp. 207-216.

[2] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, in: Proceedings of the 20th International Conference on Very Large Databases, Santiago, Chile, 1994, pp. 487-499.

[3] E. Batard, M. Lefebvre, G. Ghislain Aubin, N. Caroff, S. Corvec, "High prevalence of cross-resistance to fluoroquinolone and cotrimoxazole in tetracyclineresistant Escherichia coli human clinical isolates", Journal of Chemotherapy, 2016, DOI: 10.1179/1973947815Y.0000000038.

[4] C. L. Cazer, M. A. Al-Mamun, K. Kaniyamattam, W. J. Love, J. G. Booth, C. Lanzas, Y. T. Gröhn, "Shared multidrug resistance patterns in chickenassociated Escherichia coli identified by association rule mining", Frontiers in Microbiology, 10, 2019, 687.

[5] H. H. Chang, T. Cohen, Y. H. Grad, W. P. Hanage, T. F. O'Brien, M. Lipsitch, "Origin and proliferation of multiple-drug resistance in bacterial pathogens", Microbiol. Mol. Biol., 79, 2015, pp. 101-116.

[6] A. Comin, A. Jeremiasson, G. Kratzer, L. Keeling, "Revealing the structure of the associations between housing system, facilities, management and welfare of commercial laying hens using additive Bayesian networks", Preventive veterinary medicine, **164**, 2019, pp. 23-32.

[7] P. Congdon, Bayesian Statistical Modelling, vol. 704, John Wiley & Sons, 2007.

[8] A. Djebbari, J. Quackenbush. "Seeded Bayesian Networks: constructing genetic networks from microarray data.", BMC systems biology, **2**, 2008, pp. 1-13.

[9] S. Hartnack, T. Odoch, G. Kratzer, R. Furrer, Y. Wasteson, T. M. L'Abée-Lund, E. Skjerve, "Additive Bayesian networks for antimicrobial resistance and potential risk factors in non-typhoidal Salmonella isolates from layer hens in Uganda", BMC Veterinary Research, **15**, 212, 2019, pp. 1-9.

[10] S. Imoto, T. Higuchi, T. Goto, K. Tashiro, S. Kuhara, S. Miyano, "Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks", Proc IEEE Comput. Soc. Bioinform. Conf. 2003, **2**, pp. 104-113

[11] G. Kratzer, R. Furrer, mcmcabn: a structural MCMC sampler for DAGs learned from observed systemic datasets. R package version 0.3, 2019, https://CRAN.R-project.org/package=mcmcabn.

[12] G. Kratzer, M. Pittavino, F. I. Lewis, R. Furrer, abn: an R package for modelling multivariate data using additive Bayesian networks. R package version 2.2, 2019, https://CRAN.R-project.org/package=abn.

[13] G. Kratzer, F. I. Lewis, B. Willi, M. L. Meli, F. S. Boretti, R. Hofmann-Lehmann, P. Torgerson, R. Furrer, S. Hartnack, "Bayesian network modeling applied to feline calivicius infection among cats in Switzerland", Frontiers in Veterinary Science, **7**, 2020, pp. 1-16.

[14] F. I. Lewis, F. Brülisauer, G. J. Gunn, "Structure discovery in Bayesian networks: an analytical tool for analysing complex animal health data", Preventive Veterinary Medicine, **100**, 2011, pp. 109-115.

[15] F. I. Lewis, B. J. J. McCormick, "Revealing the complexity of health determinants in resource-poor settings", American Journal of Epidemiology, **176**, 2012, pp. 1051-1059.

[16] L. I. Lopera González, A. Derungs, O. Amft, "A Bayesian approach to rule mining", 2019, arXiv:1912.06432.

[17] A. Ludwig, P. Berthiaume, P. Boerlin, S. Gow, D. Léger, F. I. Lewis, "Identifying associations in Escherichia coli antimicrobial resistance patterns using additive Bayesian networks", Preventive Veterinary Medicine, **110**, 2013, pp. 64-75.

[18] National Antimicrobial Resistance Monitoring System (NARMS) Now: Human Data. Atlanta, Georgia: U.S. Department of Health and Human Services, CDC. 09/18/2019. https://www.cdc.gov/narmsnow Accessed 7/19/2019.

[19] D.A. Tadesse, S. Zhao, E. Tong, S. Ayers, A. Singh, M.J. Bartholomew et.al., "Antimicrobial drug resistance in Escherichia coli from humans and food animals", United States,1950–2002, Emerging Infectious Diseases, 18 (5), 2012, pp. 741-749.

[20] A. V. Werhli, D. Husmeier, "Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge", Statistical Applications in Genetics and Molecular Biology, **6**, 2007, article 15, pp. 1-45.

[21] C. Zaharia, R. Muresan, R. Moleriu, D. Zaharie, "Analysis of association measures used to discover antimicrobial resistance patterns", In: 2019 E-Health and Bioengineering Conference (EHB). IEEE, 2019, pp. 1-4.