Prediction of antibiotic resistance in Escherichia coli from large-scale pan-genome data

Mannika Garg, Raj Kamal Yadav, Rajat Bansal, Shailja Bhatt
Indraprastha Institute of Information Technology, New Delhi
Course: BIO 543 - Big Data Mining in Healthcare

Abstract— **With the advent of Artificial Intelligence, Healthcare has been blessed with various alternative diagnostic techniques[1]. Faster computation, low cost, and easy accessibility are some of the key features of computational diagnosis. Clinical methods fail to provide us with quicker and cost-efficient ways to recognize illness.**

Bacterias are evolving at a rapid pace and are growing immune toward drugs. In such circumstances, it is essential to have a quick detection mechanism for the presence or absence of drug-resistant bacteria. The conventional clinical procedure takes a day or two and can only detect the genes which are already recognized to induce drug-resistant behavior in bacterias. Therefore, the clinical method fails to identify drug-resistant bacteria with no prior information on the gene that causes resistance. Clinical practices are rule-based techniques[2] to identify resistance and hence have a limited scope and expect state-of-the-art research and development. In this paper, we aim to achieve the best machine learning model for the detection of drug-resistance[3] to 11 antibiotic compounds in 1936 Escherichia coli [4] stain from large-scale pan-genome data[5]. We have compared four models, random forest; gradient boosted decision tree, X gradient boosted decision tree, and deep neural network with the conventional rule-based approach.

Out of all the models, the gradient boosted decision tree offered the best average accuracy for all anti-biotic components (92.13%). The rule-based approach offered 74.73% accuracy, which is the lowest among all the other models.

The dataset can be split into three different components i.e., accessory genes, population structure, and year of isolation. Accessory genes alone offered decent results, but the inclusion of population structure and year of isolation followed better results.

Our research successfully improved the detection accuracy for drug-resistance by a noteworthy amount (from 74 to 92%). The study can be enhanced further by adding more data samples from different geographical locations to the dataset to incorporate diversity.

I. Introduction

With more bacterial strains becoming resistant to antibiotics, there is a significant threat of potentially incurable outbreaks. There is a need of accurate, cheap and fast pre-clinical diagnostic methods for the detection of antibiotic resistance in order to control the spread of resistant strains.

The existing methods include culture-based laboratory tests and rule-based models. The rule-based models are genetic tests which can be used to detect known resistance genes. They are an affordable solution and can accurately identify the antibiotic resistance owing to the higher availability of genome sequencing data from clinical strains[6].

The traditional culture-based laboratory tests are costly and time-consuming[7]. To overcome these limitations, genetic approaches were devised. However, these approaches can only be used to detect resistance caused by known pathogens with well-defined resistance mechanisms.

The Machine learning-based approaches cater to all the limitations from existing methods and provide a robust prediction mechanism even on unknown pathogens[8]. In this paper, we are using various machine learning models for predicting the resistance in whole-genome sequences of 1936 E. coli strains from 11 compounds across four classes of antibiotics [9][10]. These models are

then compared with the conventional rule based approach.

II. Material and Methods

A. Dataset

1936 E. coli strains were collected from various sources (Large scale clinical and environmental E.coli collections and few ecological niches). Antimicrobial susceptibility test[11][12][13] was executed on every genome sequence for each of the 11 antibiotic components. The 11 antibiotic components are - AMP (ampicillin), CXM (cefuroxime), CTX (cefotaxime), CET (cephalothin), CTZ (ceftazidime), GEN (gentamicin), TBM (tobramycin), CIP (ciprofloxacin), AMC (amoxicillin-clavulanate), AMX (amoxicillin) and TMP (trimethoprim). For every antibiotic component, the strains were labeled as S or R (S for susceptible and R for resistant).

Paired-end reads[14] were obtained from the genome sequence, which was used to construct pan-genome (union of genes in all strains) for E.coli.[15][16][17]

For the population structure dataset, we used the clustering technique on the genome sequence[22]. The final dataset can be separated into three components, i.e. Accessory gene matrix (G), population structure (P), and the year of isolation (Y). gij = 1 if the jth strain contains the ith gene (from pan-genome) and gij = 0 if the jth strain does not contain the ith gene. pij = k if jth gene belongs to kth cluster when the total number of clusters is i. yi is the year of isolation of ith strain. The total number of features is equal to 17200 for GY and 17199 for G dataset. Further, we normalized the feature vector with mean = 0 and variance = 1.

We have split the dataset in the ratio 2:8 for testing and training, respectively. We have used 5-fold cross-validation to fine-tune the best hyper parameters of the proposed models. We have tried and tested all out models on the following datasets - G,GSY,GY.

For cleaning our data set, we have eliminated feature vectors based on the correlation. We have set the threshold as 0.95 and removed all the redundant feature vectors satisfying the criteria[18]. Further, we have employed PCA and LDA to compare results derived from different models and datasets.

B. Models

1) Random Forest Classifier: We have used the "RandomForestClassifier" and "VotingClassifier" implementation from the Scikit-learn library[19]. In our model, the voting classifier consists of 3 Randomforest classifiers. We have used hard voting criteria. The number of trees in the forest (n_estimators) is 1000,1200 and 1500, respectively. We have kept oob_score as True, and all the other parameters are set as default.

2) Gradient boosted decision tree: In this model, we have used a voting estimator (with hard voting) consisting of 3 Gradient boosting classifiers, with learning rate as 0.1, from the Scikit-learn library. The number of boosting stages (n_estimator) used is 310 ,320 and 300, respectively. max_depth is 10, the minimum number of samples required to split the internal node is 2, the minimum number of samples needed to be a leaf is five, and the subsample is tuned as 0.81 for all the three classifiers.

3) Deep neural network: We have used different parameters for each of the antibiotic compounds(Table 1).Activation function used are relu and softmax. Softmax is used in the last layer only.

Value 1	Layers
AMC	(200,100)
AMP	(400,150,150)
AMX	(400,200,200,200,200,200,200)
CET	(600,150,150,150)
CIP	(400,150)
CTX	(600,150,150,150)
CTZ	(200,150,150,150,150,150)
CXM	(600,100)
GEN	(600,100)
TBM	(400,150,150,150,150,150)
TMP	(600,150,150,150,150,150)

TABLE I: Deep neural Network

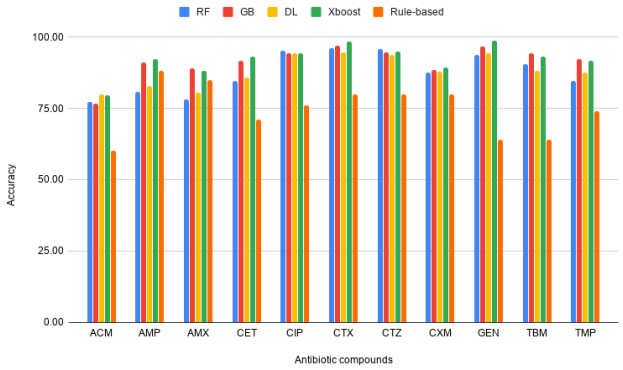


Fig. 1. Accuracy Comparison

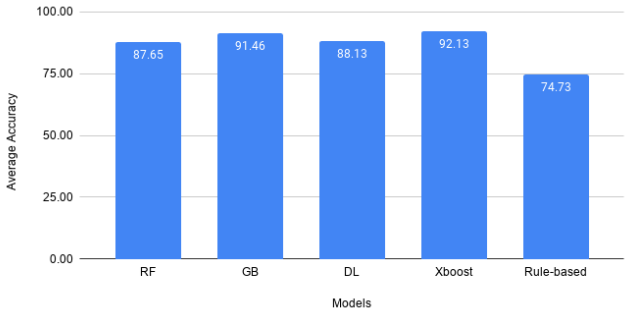


Fig. 2. Average Accuracy Comparison

4) X Gradient Boosting: In this model, we have used learning rate 0.1 and n_estimators as 500 for all the drugs.

III. Results

The big picture of this entire study can be concluded as accuracy achieved by the models proposed by us is better than the accuracy of detection by the traditional approach (Fig 1,2). The best results were obtained from GY dataset for Random forest model, Gradient boosting model and X Gradient boosting model, and from dataset G for Deep Learning model.

Accuracy is the most intuitive metric to identify the best model, but it is suitable for a uniform dataset with the same number of positives and negatives. In our training dataset, the number of susceptible and resistant strains is not equal, and thereby, we will use metrics like precision, recall, and F1 score. Fig.3 and Fig.4 demonstrate the F1 score for susceptibility and resistance, respectively.

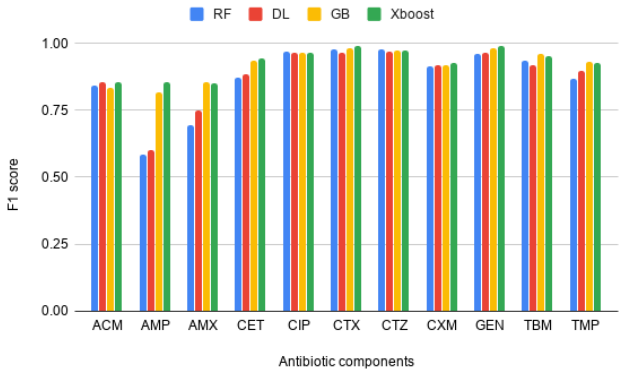


Fig. 3. F1 Score for Susceptibility

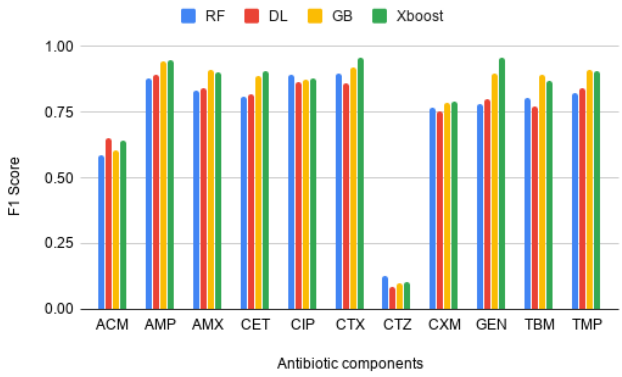


Fig. 4. F1 Score for Resistance

The F1 score for resistance from CTZ is low (Fig. 4), so we can say that our models are not suitable for the detection in the case of CTZ.
From table II, it is evident that X Gradient boosting is the best model for detection of drug resistant bacteria because the average F1 score for Resistance and Susceptibility is the highest Gradient boosting.

R or S	**RF**	**DL**	**GB**	**Xboost**
Resistance	0.74	0.74	0.79	0.81
Susceptibility	0.87	0.88	0.92	0.93

TABLE II: Average F1 Score

Table III, Table IV, Table V and Table VI is the

Antibiotic comp	TN	FP	FN	TP	Precision(S)	Precision(R)	Recall(S)	Recall(R)	F1 Score(S)	F1 Score(R)	Total S	Total R	Accuracy
ACM	204	23	54	55	0.79	0.71	0.90	0.50	0.84	0.59	227	109	77.08
AMP	21	19	11	106	0.66	0.85	0.53	0.91	0.58	0.88	40	117	80.89
AMX	54	30	18	117	0.75	0.80	0.64	0.87	0.69	0.83	84	135	78.08
CET	83	12	12	50	0.87	0.81	0.87	0.81	0.87	0.81	95	62	84.71
CIP	283	3	15	75	0.95	0.96	0.99	0.83	0.97	0.89	286	90	95.21
CTX	286	4	10	61	0.97	0.94	0.99	0.86	0.98	0.90	290	71	96.12
CTZ	321	0	14	1	0.96	1.00	1.00	0.07	0.98	0.13	321	15	95.83
CXM	252	12	35	77	0.88	0.87	0.95	0.69	0.91	0.77	264	112	87.50
GEN	309	4	20	43	0.94	0.91	0.99	0.68	0.96	0.78	313	63	93.62
TBM	111	4	11	31	0.91	0.89	0.97	0.74	0.94	0.81	115	42	90.45
TMP	78	10	14	55	0.85	0.85	0.89	0.80	0.87	0.82	88	69	84.71

TABLE III: Random Forest Analysis Table

Antibiotic comp	TN	FP	FN	TP	Precision(S)	Precision(R)	Recall(S)	Recall(R)	F1 Score(S)	F1 Score(R)	Total S	Total R	Accuracy
ACM	204	23	45	64	0.82	0.74	0.90	0.59	0.86	0.65	227	109	79.76
AMP	22	22	7	118	0.76	0.84	0.50	0.94	0.60	0.89	44	125	82.84
AMX	64	20	23	112	0.74	0.85	0.76	0.83	0.75	0.84	84	135	80.37
CET	91	5	19	54	0.83	0.92	0.95	0.74	0.88	0.82	96	73	85.80
CIP	294	4	18	71	0.94	0.95	0.99	0.80	0.96	0.87	298	89	94.32
CTX	290	5	15	62	0.95	0.93	0.98	0.81	0.97	0.86	295	77	94.62
CTZ	314	7	14	1	0.96	0.13	0.98	0.07	0.97	0.09	321	15	93.75
CXM	269	10	37	71	0.88	0.88	0.96	0.66	0.92	0.75	279	108	87.86
GEN	321	6	16	44	0.95	0.88	0.98	0.73	0.97	0.80	327	60	94.32
TBM	115	3	17	34	0.87	0.92	0.97	0.67	0.92	0.77	118	51	88.17
TMP	92	5	16	56	0.85	0.92	0.95	0.78	0.90	0.84	97	72	87.57

TABLE IV: Deep Learning Analysis Table

Antibiotic comp	TN	FP	FN	TP	Precision(S)	Precision(R)	Recall(S)	Recall(R)	F1 Score(S)	F1 Score(R)	Total S	Total R	Accuracy
ACM	197	30	49	60	0.80	0.67	0.87	0.55	0.83	0.60	227	109	76.49
AMP	31	9	5	112	0.86	0.93	0.78	0.96	0.82	0.94	40	117	91.08
AMX	70	14	10	125	0.88	0.90	0.83	0.93	0.85	0.91	84	135	89.04
CET	93	2	11	51	0.89	0.96	0.98	0.82	0.93	0.89	95	62	91.72
CIP	283	3	18	72	0.94	0.96	0.99	0.80	0.96	0.87	286	90	94.41
CTX	286	4	7	64	0.98	0.94	0.99	0.90	0.98	0.92	290	71	96.95
CTZ	317	4	14	1	0.96	0.20	0.99	0.07	0.97	0.10	321	15	94.64
CXM	251	13	31	81	0.89	0.86	0.95	0.72	0.92	0.79	264	112	88.30
GEN	313	0	12	51	0.96	1.00	1.00	0.81	0.98	0.89	313	63	96.81
TBM	111	4	5	37	0.96	0.90	0.97	0.88	0.96	0.89	115	42	94.27
TMP	84	4	8	61	0.91	0.94	0.95	0.88	0.93	0.91	88	69	92.36

TABLE V: Gradient Boosting Analysis Table

Antibiotic comp	TN	FP	FN	TP	Precision(S)	Precision(R)	Recall(S)	Recall(R)	F1 Score(S)	F1 Score(R)	Total S	Total R	Accuracy
ACM	205	22	47	62	0.81	0.74	0.90	0.57	0.86	0.64	227	109	79.46
AMP	36	4	8	109	0.82	0.96	0.90	0.93	0.86	0.95	40	117	92.36
AMX	75	9	17	118	0.82	0.93	0.89	0.87	0.85	0.90	84	135	88.13
CET	92	3	8	54	0.92	0.95	0.97	0.87	0.94	0.91	95	62	92.99
CIP	278	8	13	77	0.96	0.91	0.97	0.86	0.96	0.88	286	90	94.41
CTX	288	2	4	67	0.99	0.97	0.99	0.94	0.99	0.96	290	71	98.34
CTZ	318	3	14	1	0.96	0.25	0.99	0.07	0.97	0.11	321	15	94.94
CXM	260	4	36	76	0.88	0.95	0.98	0.68	0.93	0.79	264	112	89.36
GEN	313	0	5	58	0.98	1.00	1.00	0.92	0.99	0.96	313	63	98.67
TBM	110	5	6	36	0.95	0.88	0.96	0.86	0.95	0.87	115	42	92.99
TMP	82	6	7	62	0.92	0.91	0.93	0.90	0.93	0.91	88	69	91.72

TABLE VI: X Gradient Boosting Analysis Table

final summary of the four models we proposed. On closer analysis of F1 scores for each drug, we learn that X gradient boosting outperforms all the other models for 6/11 drugs (AMP, CET, CTX, CXM, GEN, TMP) whereas gradient boosting performs best for 2 out of 11 drugs(AMX, TBM). Deep learning model performs best for 1/11 drugs (AMC) and Random forest performs best for CIP and CTZ.

Figure 5 illustrates the resemblance between our method and the-state-of-art. Our approach has outperformed the state-of-the-art model for drug CTX by 0.25%.

IV. Discussion

In the implementation, we have compared the rule-based model (clinical methodology) with the four proposed models, i.e., Random Forest classifier, Gradient Boosting, X Gradient Boosting

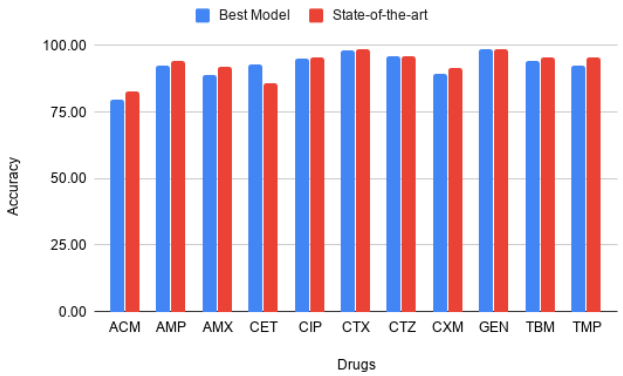
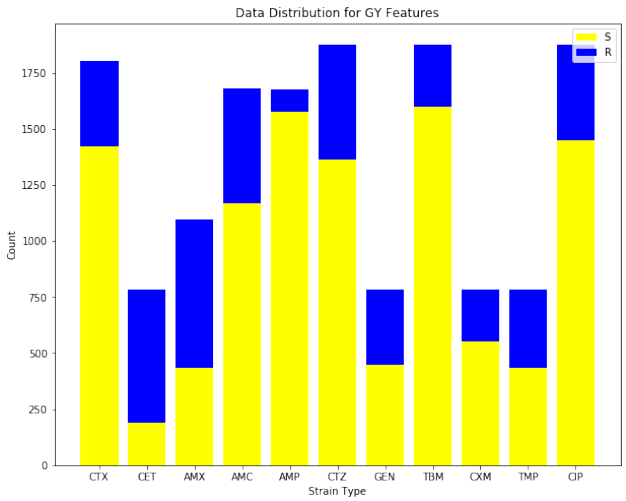


Fig. 5. Accuracy Comparison between our best model and state-of-the-art

and deep learning. To compare the models, we have used the Accuracy, Precision, Recall, and F1 score. We have used 5-fold cross-validation to select the best hyperparameters for the following model. The data consists of 3 sections, accessory genes (pan-genome), population structure, and year of isolation. Accessory genes are sufficient to achieve decent accuracy, but the addition of year of isolation to the feature vector has improved the accuracy by an average of 1.3% for Random forest and gradient boosting models. Population structure matrix as input feature vector has offered an accuracy of 72%.Population structure and Accessory gene data have been derived from the same genomic sequences and hence have a high correlation (0.8). We have removed all the redundant correlated data from the feature vector. The utilization of PCA on the feature vector has further strengthened the model's ability to predict by an average of 2.5%. We also observed that LDA + deep learning increased the detection accuracy by a considerable amount for a few antibiotic substances but, on average, demonstrated poor results[20].

The data collected from large-scale clinical and environmental E.coli collection is diverse, but we felt that the data from ecological niches were underrepresented (255 out of 1936). The study is a prototype and can be implemented worldwide with the help of more data from across the globe. The data was labeled in a binary fashion, which is far from reality because the intensity with which a drug affects a bacteria can be scaled.The difference in F1 score and Accuracy is the result of non-uniformity in the dataset i.e. high disparity in number of susceptible strain and number of resistant stain(Fig 6). It is, therefore, essential to have a better representation of the data for better results.

In this paper, we have established that machine learning approaches are better than the clinical procedure to detect drug resistance. X Gradient boosting offered the best accuracy among all the other models. It is essential to mention that all drug resistance mechanisms are not genomic and cannot be predicted with the given feature vector. To accommodate all such drugs, we need to expand the data and then use machine learning techniques for the detection of drug resistance.



V. Contribution of Author

All the authors devised the idea and theory. Shailja Bhatt and Mannika Garg performed data processing, analyzed results, wrote the manuscript. The model implementations and experiments were carried out by Rajat Bansal and Raj Kamal Yadav. All the authors contributed equally to the discussion and helped shape the study.

References

[1] Wikipedia contributors. Artificial intelligence in healthcare. Wikipedia, The Free Encyclopedia. May 12, 2020, 10:30 UTC. Available at:

https://en.wikipedia.org/w/index.php?title=Artificial_intelligence_in_healthcareoldid=956255199. Accessed May 20, 2020.

[2] Fluit AC, Visser MR, Schmitz FJ. Molecular detection of antimicrobial resistance. Clin Microbiol Rev. 2001;14(4):836871. doi:10.1128/CMR.14.4.836-871.2001

[3] "Machine Learning takes on Antibiotic resistance", https://www.quantamagazine.org/machine-learning-takes-on-antibiotic-resistance-20200309/

[4] E.Coli, https://www.who.int/news-room/fact-sheets/detail/e-coli

[5] Wikipedia contributors. Pan-genome. Wikipedia, The Free Encyclopedia. April 3, 2020, 21:55 UTC. Available at: https://en.wikipedia.org/w/index.php?title=Pan-genomeoldid=948949318. Accessed May 20, 2020.

[6] Aun E, Brauer A, Kisand V, Tenson T, Remm M. A k-mer-based method for the identification of phenotype-associated genomic biomarkers and predicting phenotypes of sequenced bacteria. PLoS Comput Biol. 2018;14(10):e1006434. Epub 2018/10/23. pmid:30346947; PubMed Central PMCID: PMCPMC6211763.

[7] Mitsakakis K, Kaman WE, Elshout G, Specht M, Hays JP. Challenges in identifying antibiotic resistance targets for point-of-care diagnostics in general practice. Future Microbiol. 2018;13(10):11571164. doi:10.2217/fmb-2018-0084

[8] "Machine Learning leads to speedy screening of drug-resistant microbes", 2019. https://www.nature.com/articles/d41586-019-03668-0

[9] Kallonen T, Brodrick HJ, Harris SR, Corander J, Brown NM, Martin V, et al. Systematic longitudinal survey of invasive Escherichia coli in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131. Genome Res. 2017. pmid:28720578; PubMed Central PMCID: PMCPMC5538559.

[10] Runcharoen C, Raven KE, Reuter S, Kallonen T, Paksanont S, Thammachote J, et al. Whole genome sequencing of ESBL-producing Escherichia coli isolated from patients, farm waste and canals in Thailand. Genome Med. 2017;9(1):81. pmid:28877757; PubMed Central PMCID: PMCPMC5588602.

[11] Ericsson JM, Sherris JC, 1971. Antibiotic sensitivity testing: report of an international collaborative study

[12] Balows A., 1972. Current techniques for antibiotic susceptibility testing, Springfield

[13] L. Barth Reller, Melvin Weinstein, James H. Jorgensen, Mary Jane Ferraro, Antimicrobial Susceptibility Testing: A Review of General Principles and Contemporary Practices, Clinical Infectious Diseases, Volume 49, Issue 11, 1 December 2009, Pages 17491755, https://doi.org/10.1086/647952

[14] "What are pair-end reads?", https://thesequencingcenter.com/knowledge-base/what-are-paired-end-reads/

[15] Page AJ, De Silva N, Hunt M, Quail MA, Parkhill J, Harris SR, et al. Robust high-throughput prokaryote de novo assembly and improvement pipeline for Illumina data. Microb Genom. 2016;2(8):e000083. pmid:28348874; PubMed Central PMCID: PMCPMC5320598.

[16] Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):20689. pmid:24642063.

[17] Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31(22):36913. pmid:26198102; PubMed Central

PMCID: PMCPMC4817141.

[18] ”Applying Filter method in Python for Feature Selection”, https://stackabuse.com/applying-filter-methods-in-python-for-feature-selection/

[19] Pedregosa et al., Scikit-learn: Machine Learning in Python, JMLR, 2011, 12, pp. 28252830,.

[20] Jones William, Alasoo Kaur, Fishman Dmytro, Parts Leopold, Computational biology: deep learning, 2017, , Emerging Topics in Life Sciences

[21] Definition of Antimicrobial Susceptibility Testing (AST), https://www.synbiosis.com/support/glossary/definition/?term=Antimicrobial%20Susceptibility%20Testing%20(AST)

[22] Jombart T. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics. 2008;24(11):14035. pmid:18397895.