BenchmarkDR

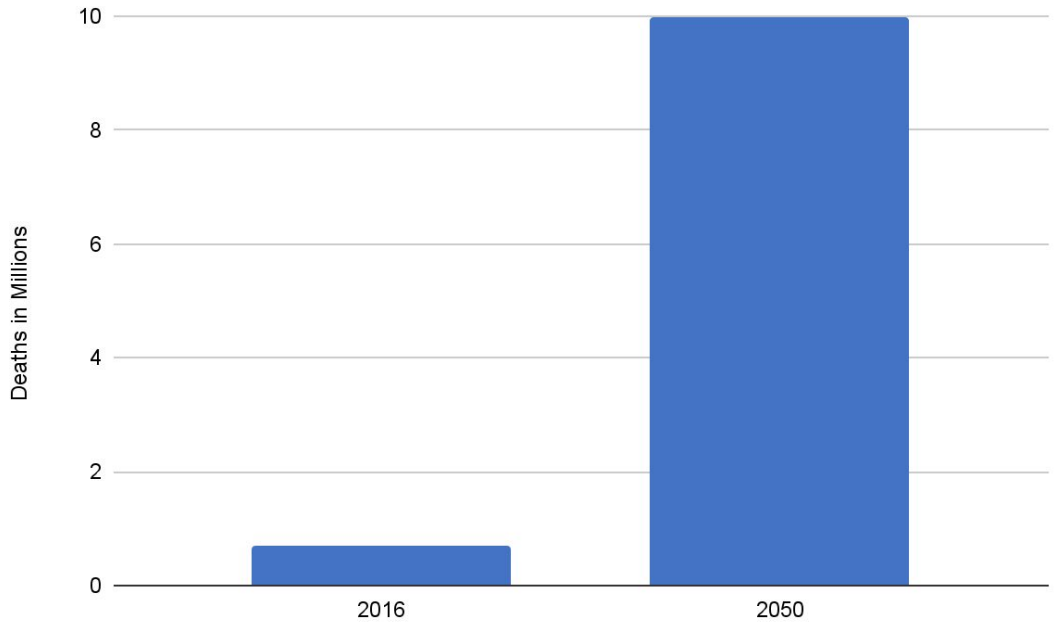
A modular and expandable benchmarking pipeline for machine learning based antimicrobial resistance prediction

Niklas Stotzem*, Fernando Guntoro*, Leonid Chindelevitch*

*Imperial College London

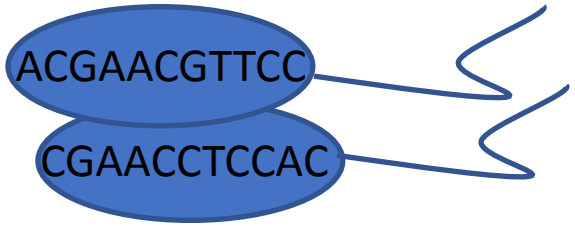
Multi-drug-resistant pathogenic bacteria are an increasing global threat

Deaths linked to multi-drug-resistant (MDR) microbials (J. O'Neill, 2016)



Machine Learning (ML) is a useful tool to fight pathogenic drug resistant bacteria

Genomic NGS Data



Drug Susceptibility Test Data

Bacterial Isolate	Antibiotic 1	…
1	Resistant	…
2		
…		

Bacterial Isolate	Antibiotic 1	…
1	MIC X	…
2	MIC Y	…
…	…	…

ML Methods for Classification and Regression

All Models
Clinical Diagnostics

Interpretable Models
Identification of Biomarkers
Research of resistance causing mechanisms

Different features of the genome as well as different ML methods have been used by researchers

Representations

Genes

…ATCAAATCCGTTTCAAGGTCCCTTGCCAACCGGTTGGAACG

Single Nucleotide Polymorphisms (SNPs)

…ATCAAATCCGTTTCAAGGTCCTTGCCAACCGGTTGGAACG

A G

K-MERs

…ATCAAATCCGTTTCAAGGTCCCTTGCCAACCGGTTGGAACG

ATCA

TCAA

……

ML Methods

Logistic Regression

Linear Regression

Random Forests

Support Vector Machines

…

New Ones

How to compare methods performance on different representations of different data?

We built an end-to-end pipeline to allow users to compare representations and methods

What does the pipeline have under the hood?

- **Representations**
  - Tools based on benchmark papers, popularity and ease of use

- **ML Methods**
  - 17 for Classification (Scikit-Learn (Pedregosa, F. et al., 2011) & INGOT-DR (Zabeti, H. et al., 2021))
  - 9 for Regression

- **Evaluation**
  - Time and Prediction Performance

Configuration can be adapted at different levels of detail

```yaml
PATH_DATA: # Path to folder in which pneumonia_mic folder is located as string

BACTERIA: ["pneumonia_mic"]

OUTPUT_DIR:  # Path to folder in which all further downstream files will be saved as string

REPRESENTATION: ["gene_presence_absence", "snps", "kmer"]

MODE: "MIC" # "Classification" or "MIC"

DRUGS: ['Tobramycin',
 'Cefuroxime sodium',
 'Ciprofloxacin',
 'Gentamicin',
 'Ampicillin',
]

METHODS: ["sklearn_LinR", "sklearn_LinR_l1", "sklearn_LinR_l2", "sklearn_LinR_elasticnet"]

OPTIMIZATION: "GridSearchCV" # "GridSearchCV", "RandomizedSearchCV" and "None"
```

```yaml
  sklearn_ADBC:
    module: sklearn.ensemble
    model: AdaBoostClassifier
    params:
      n_estimators: 50
      learning_rate: 0.0001
    cv:
      n_estimators: [50, 100, 500, 1000]
      learning_rate: [0.0001, 0.001, 0.01, 0.1, 1.0]

  sklearn_GBTC:
    module: sklearn.ensemble
    model: GradientBoostingClassifier
    params:
      max_features: 'auto'
      n_iter_no_change: 20
    cv:
      learning_rate: [0.001, 0.01, 0.1, 1]
      n_estimators: [100, 300, 500, 1000]
      min_samples_split: [2, 5, 10, 15]
      max_depth: [5, 10, 15, 30]
```

A dashboard prototype provides an convenient overview of results and further details

DRUG	MODEL	TIME	ACCURACY	BALANCED_ACCURACY	F1_SCORE	ROC_AUC	SENSITIVITY	SPECIFICITY
amikacin	GBT	44466.423	0.868	0.83	0.773	0.83	0.73	0.929
capreomycin	ADB	677.97	0.835	0.776	0.697	0.776	0.623	0.928
capreomycin	SVM_l2	252.93	0.81	0.776	0.689	0.776	0.689	0.863
ciprofloxacin	LR_l2	11.58	0.889	0.795	0.615	0.795	0.667	0.923
ethambutol	SVM_l2	892.432	0.848	0.821	0.726	0.821	0.764	0.878
ethionamide	LR_l1	16.649	0.743	0.724	0.642	0.724	0.66	0.788
isoniazid	ADB	2784.837	0.864	0.856	0.836	0.856	0.777	0.935

Conclusions & Future Outlook

- Easily useable and extensible end-to-end pipeline to benchmark 26 different ML methods on genomic data represented in 3 ways

- Applicable for other microbial phenotype predictions

Future Work

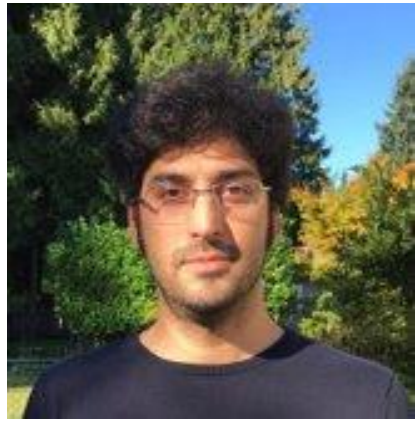
- Further Extensions (Methods, Representations)
- Provision of comprehensive gold standard datasets
- Adding explainability approaches, e.g. SHAP
- Include long-read sequencing data
- Grow the tool with the community

Acknowledgements

Repository: https://github.com/WGS-TB/BenchmarkDR

- Fernando Guntoro
- Dr. Leonid Chindelevitch (Supervision)
- Dr. John Lees (Supervision)
- Dr. Hooman Zabeti (Supervision)



Key References

Guntoro, F. (2021), Benchmarking antimicrobial resistance prediction with an automated pipeline for machine learning models using BenchmarkDR [Master's Thesis, Imperial College London].

J. O'Neill (2016). Review on Antimicrobial Resistance. *Tackling Drugresistant Infections Globally: Final Report and Recommendations*. Review on Antimicrobial Resistance, 2016.

Pedregosa, F. et al. (2011), *Scikit-learn: Machine Learning in Python*, Journal of Machine Learning Research, 12.

Stotzem, N. (2021), An End-to-end Machine Learning Pipeline for Drug Resistance Prediction in Bacteria [Master's Thesis, Imperial College London].

Zabeti, H., Dexter, N., Safari, A. H., Sedaghat, N., Libbrecht, M., & Chindelevitch, L. (2021). *INGOT-DR: an interpretable classifier for predicting drug resistance in M. tuberculosis*. Algorithms for Molecular Biology, 16(1), 17.

Images

https://theconversation.com/antibiotic-resistance-new-discovery-could-change-the-future-of-treatment-131262

https://github.com/snakemake

Backup

We selected the tools used to create the representations based on published benchmarks, popularity and ease of use

Representation	**Tools**
Gene Presence/ Absence	SPAdes v3.15.2 (Assembly), Prokka v1.13.4 (Annotation)
Single Nucleotide Polymorphisms (Requires Reference Genome)	BWA v0.1.17 (Alignment), Samtools v1.12 (Sorting), Picard v2.25.6 (Duplicate Removal), VarScan v2.4.4 (SNP calling)
K-mers	KMC v3.1.2.rc1

The pipeline includes - so far - a variety of standard ML methods and INGOT-DR

Binary Classification (Resistant/ Susceptible)	Regression (MIC)
Logistic Regression* Support Vector Machine Classification* Decision Trees Random Forests Extremely Randomized Trees AdaBoost Decision Tree Classifier Gradient Boosted Decision Trees Stochastic Gradient Descent Classifier* K-Nearest Neighbours Gaussian/ Complement Naive Bayes INterpretable GrOup Testing for Drug Resistance (INGOT-DR) (Zabeti, H. et al., 2021)	Linear Regression* Support Vector Machine Regression Decision Tree Regressor Random Forest Regressor Gradient Boosted Trees Regressor AdaBoost Decision Tree Regressor

* incl. variations with l1 and l2 regularization

Several evaluation metrics are provided

Classification	**Regression**
Accuracy, Balanced Accuracy, F1-Score, AUC, Sensitivity, Specificity	Mean Squared Error, Mean Squared Log Error, Coefficient of Determination