Deliverable 3.2

Prediction algorithm for antimicrobial resistance markers in sequence data

Version:

Due: Month 24

Completed: Month 24

Authors: C. Schultsz; S. Matamoros; Basil Britto Xavier; Surbhi Malhotra

Contributing Partners: AMC, UA

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 643476.

Contents

Deliverable Description

Bacterial genomic data are becoming increasingly easy to collect, since cost and data-analysis time have decreased drastically in the past years. One of the main advantages of whole genome sequence (WGS) analysis is the possibility to predict specific phenotypes based on genome content: e.g. virulence and antimicrobial resistance (AMR).

AMR in clinical isolates is determined using standardized culture-based methods (e.g., EUCAST guidelines). The resulting phenotypes are currently the gold standard, and each new method of AMR susceptibility testing will have to be compared to it.

Sequence based databases (e.g. ResFinder; CARD) contain information about genes that have been shown to encode AMR in laboratory experiments. Promising studies have shown that assembling large numbers of known AMR-encoding genes can lead to accurate prediction of the AMR characteristics of a specific isolate from its genome. It has the advantage of establishing a direct link between genotype and phenotype. However, this approach does not take the cumulative effect of several AMR encoding genes, including those encoding for efflux pumps, on a single AMR phenotype into account and is restricted to previous knowledge about the function of each genetic element. The recent discovery of the *mcr-1* gene providing resistance to colistin has shown that knowledge gaps still exist, increasing the risk of false negative results when searching for AMR genotypes. In addition, AMR encoded through inter-genic sequences or otherwise encoded regulatory mechanisms will be missed whilst no inferences can be made based on the absence of genes that are listed in the database. Finally, results are likely to be binary, i.e. the presence of absence of a gene predicts a resistant or susceptible phenotype, not allowing for potentially more subtle predictions that correlate with a more continuous measure of the AMR phenotype such as the minimum inhibitory concentration (MIC).

Recent studies have shown that pattern recognition techniques such as machine learning allow the creation of models and algorithms that can accurately relate genomic and phenotypic information. However, to our knowledge, this technique has only been used retrospectively to show an association between genomic content and AMR phenotypes.

Thus a big-data approach, using pattern-recognition on a large number of bacterial genomes could allow for an unbiased prediction of the AMR profile of a bacterial isolate without the corresponding phenotypic data. A machine learning approach could potentially be used to infer susceptibility patterns from the genomes of unknown isolates independent of the identification of specific genes or sequences that are associated with a particular phenotype through molecular biology analyses.

Also, our review on existing resistance gene databases (Xavier 2016 in JCM doi: 10.1128/JCM.02717-15. Epub 2016 Jan 27.) highlighted the pitfalls and advantages of the existing databases. None of the database have any association of phenotypic and genotypic resistance data.

Deliverable 3.2 is an algorithm capable of predicting AMR profiles directly from bacterial isolates sequence data.

For this purpose, we divided the task into subgroups and focused on specific antibiotic groups and different approaches.

1. A Machine learning approach (ML) is being utilized todevelopan algorithm in a broader fluoroquinolone and *E. coli*, *Salmonella* (AMC)
2. A hidden Markov models (hmm) database is being developed for all colistin resistance genes, plasmid and chromosomal specific, especially for Enterobacteriaceae (UA)

In order to reach this goal, several COMPARE teams (EBI/ENA, DTU, UNIBO, WIGNER) under the coordination of AMC (workbundle…leader, WP3/6) have come together to assemble the expertise necessary for the design of the algorithm, gather the data required (bacterial genomes and corresponding AMR profiles) and make it available to all partners.

As a first step towards the creation of a reliable AMR prediction algorithm, an online database was created to gather the data (WGS and AMR profiles) and make it readily available to all the partners involved in the project.

Organization of the database for ML (AMC)

The database is constituted of a private online repository hosted by the EBI/ENA, based on the ENA template for online storage and access to bacterial genome sequences. Data providers (COMPARE teams and laboratories possessing WGS and related AMR metadata) can connect to the database and upload the bacterial WGS data. The accepted format is raw fastq reads, which is the standard output from Illumina and Ion Torrent sequencers.

The corresponding phenotypic AMR metadata is also submitted via the same procedure, and the data upload is notified to the EBI. The AMR metadata is stored in a standard format, identical for all isolates and allowing its analysis in lalsirge batches.

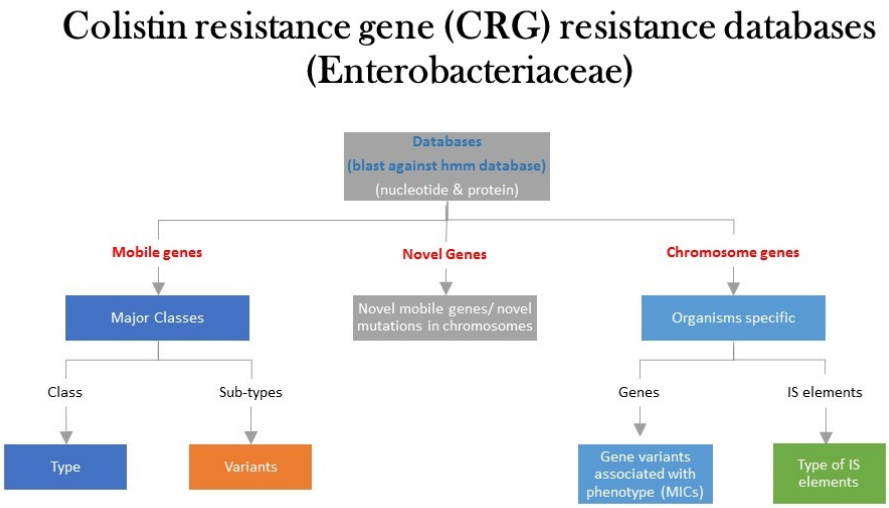
COMPARE data consumers (teams with expertise in machine learning programming) can access the database to retrieve the data and use it in the design and testing of their AMR prediction algorithm.

The database accepts genomes that have already been published (public) or not (private). A specific access to the database must be requested for data consumers and agreed upon by data providers.

HMM databases for colistin resistance genes (UA)

We have collected all resistance genes sequences from Enterobacteriaceae that have till now been shown to be involved in colistin resistance in literature, both chromosomal and plasmid-based. HMM databases of involved protein and nucleotide sequences are being built separately. Based on our literature search on colistin resistance studies we could extract some phenotypic-genotypic correlations. However this information is not complete. We would like to expand and complete this information and validate our HMM based colistin resistance gene database (Figure 1). We have in-house a large number of clinical

colistin resistant strains collected through various projects that will be tested in the laboratory (cloning, site-directed mutagenesis etc) to build pheno-genotypic correlations to add to our currently limited database.



Data submission procedure

The data sharing falls under COMPARE general agreement for collaborative projects. A list of data providers and data consumers is kept up to date to ensure their access to the database. The data cannot be shared outside of COMPARE without express consent from all interested parties. All data providers will be considered as co-authors for the publications resulting from this project.

In order to facilitate the submission of data, 2 tutorial documents have been written. They describe in details the steps to perform for the Submission of WGS and AMR data (the first page of each document is presented in annexes 1 and 2). AMC provides additional support to data providers in case they encounter unanticipated problems.

A template of AMR data submission was jointly created by AMC, DTU and EBI (see Annex III). Data providers can use this template to provide AMR data for the isolates they submit to the database. This template can be used in combination with machines such as VITEK and laboratory analysis pipeline to provide an automated output.

All information regarding data submission, including the tutorials, templates and additional validation tools that can be helpful for data providers are gathered in an online repository at the following adress:

https://github.com/EBI-COMMUNITY/compare-amr

Delays

The database was originally planned to be accessible in December 2016. However, due to delays its final version was released at the end of January 2017.

The technical setup of the database proved to be more challenging than anticipated. Adding a new feature (linkage of AMR metadata to WGS data) to the already existing ENA database was a non-trivial task that required a substantial investment in man-hour from the EBI team.

Additionally, in order to reach the objective of the deliverable of making this algorithm flexible for different organisms such as *Enterobacteriaceae*, *S. aureus*, *Enterococci* or *Mycobacterium*, the design of the database had to account for a larger number of possible bacterial species to be submitted along with complete AMR profiles. This impacted notably the design of the AMR metadata submission template.

With the database now in place, we expect on reaching the objective of 1000 *E. coli* genomes rapidly. This is the estimated size of the database necessary for an accurate comparison of the existing machine-learning models, fine tuning of their parameters and ultimately the selection of the most appropriate algorithm for this project, pre-requisite for the design of a reliable AMR prediction tool. The new estimated timeline for this deliverable is M36. This corresponds to the time required to gather the data (WGS and AMR), test the different models, evaluate their predictive power and compare the different approaches from the 3 teams engaged in the design of the algorithm.

It is noteworthy that this deliverable is a pre-requisite for deliverable 6.2: Report on WGS and NGS based detection of antimicrobial resistance in stool samples in patients and travelers which is planned for Month 50. However we estimate that the design of a robust database and the setup of a network of COMPARE teams and laboratories with the required expertise will greatly benefit Deliverable 6.2. Thus the delays observed for Deliverable 3.2 should not impact negatively the timeline for Deliverable 6.2.

Annex I. Tutorial 1: Submitting read data

Submitting read data

Our public documentation is available in: http://www.ebi.ac.uk/ena/submit/read-submission

For reporting data into the data hub, the following needs to happen:

[1] Data providers should register a submission account (Webin-NNNN) or login into an existing account here: https://www.ebi.ac.uk/ena/submit/sra/#registration

Once logged in, go to the 'New Submission' tab and click on "launch uploader". This opens a file uploader which you can use for your read submission at a later stage.

First continue with the next steps:

[2] Data providers should register projects/studies: On the 'New Submission' tab, select 'submit sequence reads and experiments' and click "next".

Either select an existing study or click on "create new study" and complete the details of your study to which the reads will be linked. Your study will appear in the list. Select it and click "next".

[3] Register samples: To register samples, a reporting standard needs to be selected. For purposes of this project, select "other checklists" and "ENA default sample checklist" and click "next".

There are 2 ways to create samples. We recommend using the 2nd option.

Option 1:
Fill out the mandatory fields only and click "next". You can now indicate how many samples you want to create according to how many read files you wish to submit. Create 1 sample per read file. Click "add" and "next".

Option 2:
Click on "Download template". The file is a tab-delimited text file (.tsv). Open excel; go to "file" then to "open"; select the data template file; it will automatically import and convert to excel format by clicking "next" and "finish". Here is an example of what that should look like:

Annex II. Tutorial 2: Submitting AMR data

Antimicrobial susceptibility testing data submission

A) Submission Procedure:

[1] Account registration:

Data providers should already have an EBI/ENA account and have submitted the read data they want to link with MICs (see tutorial on "Submitting read data").

[2] Data upload:

Create a directory called "antibiogram" on your computer and place the files containing your susceptibility data inside it.
Connect to the EBI/ENA website (http://www.ebi.ac.uk/ena). Go to the "Submit & Update" tab, then click on "Submit to ENA". Enter your account information to connect (Webin-NNNN and password).
Go to the "New submission" tab and click on "Launch uploader". You will then access the Webin Java web start EBI uploader. Enter your Webin identifier ("username") and Webin password, then connect. Browse to the "Upload Directory" containing your files ("antibiogram"). Select the files you want to upload. Tick the boxes "Overwrite" and "Upload Tree" to copy the directory and the files. Click on "Upload" to start the transfer.

If you want to use a different method or are having trouble with the Java web start application, you can use FTP (file transfer protocol) or Aspera to upload the files. Note that if you use an FTP client such as FileZilla you will have to create manually a directory called "antibiogram" on the EBI Remote site, in which you will then transfer the susceptibility testing data files.

More information and help on the file transfer can be found at this address:
https://www.ebi.ac.uk/ena/submit/uploading-data-files

[3] Data upload confirmation:

Please email to datasub@ebi.ac.uk and let us know about your upload.
Please provide "COMPARE: Antibiogram" in your email subject. And in the body of your email please provide the name of the submitted files.

Annex III. AMR data submission template

bioSample_ID	species	antibiotic_name	ast_standard	breakpoint_version	laboratory_typing_method	measurement	measurement_units	measurement_sign	resistance_phenotype	platform
SAMEA4350781	E. coli	ciprofloxacine	EUCAST	2015	Microbroth dilution	16	mg/L	=	resistant	Vitek
SAMEA4350782	E. coli	ciprofloxacine	EUCAST	2015	Microbroth dilution	16	mg/L	=	resistant	Vitek
SAMEA4350783	E. coli	ciprofloxacine	EUCAST	2015	Microbroth dilution	0.002	mg/L	<	susceptible	Vitek