



Sapienza University of Rome
Department of Biochemical Sciences “A. Rossi Fanelli”

Identification, analysis and inference of point mutations associated to drug resistance in bacteria: a lesson learnt from the resistance of *Streptococcus pneumoniae* to quinolones

David Sasah Staid

XXXIII cycle

PhD Programme in Biochemistry

Tutor
Prof. Andrea Bellelli

Supervisor
Dr. Allegra Via

Abstract

Antibiotic resistance is one of the biggest public health challenges of our time. Bacterial chemoresistance is the phenomenon whereby bacteria develop the ability to survive and multiply in the presence of an antibacterial drug; the expression of a resistant phenotype may be due to three fundamental mechanisms, including the expression of enzymes that inactivate the antibacterial drug, changes in the membrane permeability to antibiotics and the onset of point mutations causing the physical-chemical alteration of the antimicrobial targets. In recent decades, new antibiotic resistance mechanisms have emerged and are spreading globally, threatening human health and the ability to fight the most common infectious diseases.

Quinolones, a novel class of antibiotics that bind bacterial topoisomerases and inhibit cell replication, have been important in limiting the spread of penicillin- and macrolides-resistant *Streptococcus pneumoniae*. However, alarmingly, resistance to quinolones is spreading recently. Resistance is caused by the appearance of point mutations in the bacterial topoisomerase and gyrase. Some mutations are well known, but some are not and the information about known molecular mechanisms causing resistance is sparse and not systematically collected and organised. This means that it cannot be used to infer new mutations in newly sequenced bacterial genes and study how

they may affect the drug binding. The lack of structured, organized, and reusable information about point mutations associated with antibiotic resistance represents a critical issue and is a common pattern in the field.

Here, we present a structural analysis of point mutations involved in the resistance to quinolones affecting the gyrase and topoisomerase genes in *Streptococcus pneumoniae*. Results, extended to other bacterial species, have

been collected in a database, Quinores3D db, and can now be used – through a web server, Quinores3D finder - to analyze both known and yet unknown mutations occurring in bacterial topoisomerases and gyrases. The development, testing and deployment of Quinores3D db and Quinores3D finder are further results of this PhD thesis.

Furthermore, structural data about point mutations associated with antibiotic resistance were used to train, test and validate a machine learning algorithm for the inference of still unknown mutations potentially involved in bacterial resistance to quinolone. As the performance of the algorithm, measured in terms of accuracy, sensitivity and specificity, is very promising, we plan to incorporate it in the web server to allow users to predict new mutations associated with bacterial resistance to quinolones.

Index

1. Introduction	7
1.1. A brief history of antibiotics	7
1.2. Antibiotics: classes and mechanism of action	9
1.3. Quinolones and fluoroquinolones.....	12
1.3.1. Quinolones target: DNA gyrase and topoisomerase IV.....	14
1.3.2. Structure of topoisomerase II enzymes	15
1.4. The antibiotic resistance.....	20
1.5. Mechanisms of quinolone resistance	22
1.5.1. Plasmid-mediated quinolone resistance (PMQR).....	23
1.5.2. Reduced drug accumulation	23
1.5.3. Mutations within the QRDRs	24
1.6. Bioinformatics resources to fight antimicrobial resistance	25
1.6.1. The Comprehensive Antibiotic Resistance Database.....	28
1.6.2. PointFinder tool and database.....	29
1.7. Structural analysis of point mutations	30
1.9. Machine learning	33
1.9.1. Common ML algorithms employed in bioinformatics.....	35
1.9.5. Evaluation methods	41
2. Aim of the thesis.....	43
3. Materials and methods.....	45
3.1. Topoisomerase and gyrase sequences	45
3.2. Pairwise and multiple sequence alignment	46
3.3. Homology modelling protocol	47
3.4. Structural analysis.....	50
3.4.1. Quinolone binding site analysis	50
3.4.2. Electrostatic analysis	51

3.4.3.	Relative solvent accessible surface analysis.....	52
3.4.4.	Protein structure visualization and analysis	52
3.5.	Primer design.....	53
3.6.	Annotation of mutations and information retrieval.....	53
3.7.	Web server technical specification.....	55
3.8.	Data analysis and code scripting	56
3.9.	Application of information theory on sequence analysis	57
3.10.	Machine learning	57
3.10.1.	Feature encoding for the predictor model.....	58
3.10.2.	K-means clustering.....	59
3.10.3.	Training, validation and test set.....	59
3.10.5.	Metrics and statistical tests	61
4.	Results	65
4.1.	Three-dimensional modelling of topoisomeraseIV and gyrase protein structures	65
4.1.1.	Topoisomerase IV ParC subunit model: a case study for the homology modelling procedure	66
4.2.	Point mutation characterization and structural analysis	68
4.3.	Quinores3D: a web server for the structural analysis of the molecular mechanisms of resistance to quinolones.....	69
4.3.1.	Quinores3D db.....	70
4.3.2.	Quinores3D finder.....	77
4.3.3.	Quinores3D primers.	83
4.4.	Analysis of point mutations associated with quinolone resistance.....	84
4.5.	Cluster analysis of mutations associated with bacterial AR	87
4.5.1.	Cluster I_III characterization	90
4.5.1.1.	Characterization of the ParC 83 / GyrA 85 position	93
4.5.1.2.	Characterization of the ParE 435 / Gyrb 435 position.....	98
4.5.1.3.	Characterization of the ParE/GyrB 475 and ParE/ GyrB 474 positions	101

4.5.1.4.	Characterization of the ParC 78 position	106
4.5.2.	Cluster II characterization.....	108
4.5.3.	Cluster IV characterization.....	110
4.5.4.	Cluster V characterization.....	112
4.5.4.1.	Characterization of the ParC 79/ GyrA 81 position.....	117
4.6.	Semi-automatized pipeline for the identification of point mutations in <i>S. pneumoniae</i>	120
4.7.	<i>Quinores3D_pred</i> : machine learning algorithms for the prediction of point mutations	121
4.7.1.	Training, test and validation set.....	122
4.7.2.	Feature extraction.....	124
5.	Conclusions	132
6.	Bibliography	136

1. Introduction

1.1. A brief history of antibiotics

The word *antibiotic* was first used by Selman Waksman in 1941 to describe any small molecule made by a microbe that antagonizes the growth of other microbes¹. Yet, the use of antimicrobial agents can be dated back to 2000 years ago, when ancient Greeks and Egyptians were used to apply mouldy bread to treat open wounds².

The development of anti-infective drugs is widely accredited to Paul Erlich, who developed the *salvarsan*, an arsenic-based pro-drugs to treat the agent of syphilis in 1910³. Years later, Gerhard Domagk at Bayer discovered the sulfonamide prodrug *Prontosil*, which like other sulfonamides was the first effective broad spectrum antimicrobial still in use today.

The turning point came in 1928 with the discover of penicillin by Alexander Fleming, which started the golden age of natural product antibiotic discovery that peaked in the mid-1950s, especially thanks to the studies of Selman Waksman who discovered numerous antibiotics made by soil-dwelling actinomycetes, including neomycin and streptomycin².

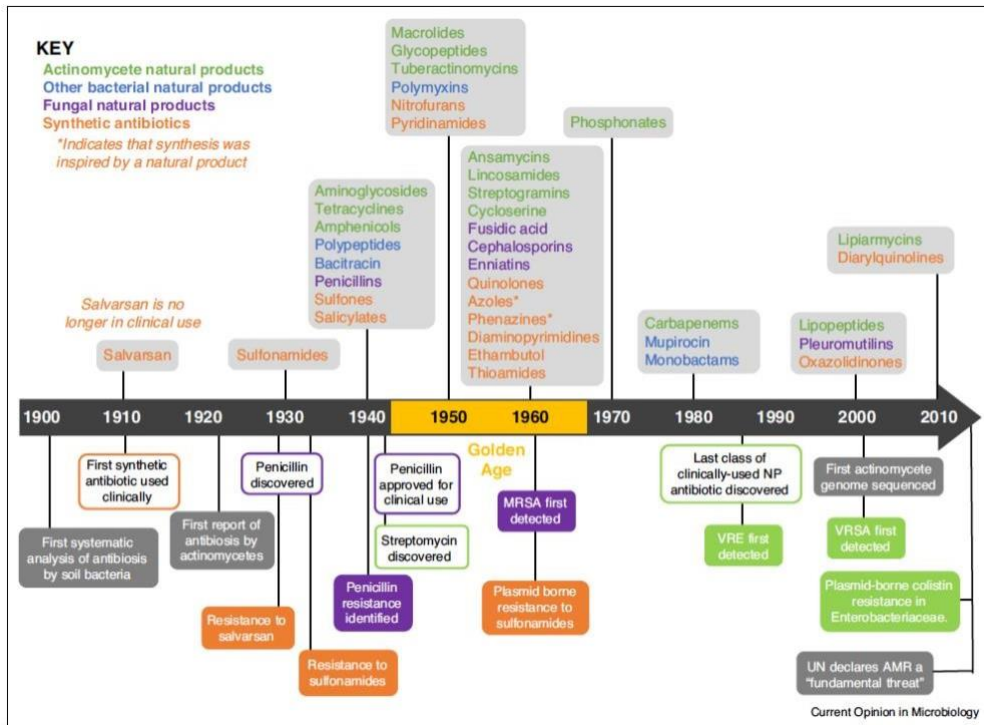


Figure 1.1. Timeline showing the decade new classes of antibiotic reached the clinic².

In just over 100 years antibiotics have drastically changed modern medicine and extended the average human lifespan by 23 years, but then, a gradual decline in antibiotic discovery and development and the evolution of drug resistance in many human pathogens has led to the current antimicrobial resistance crisis².

1.2. Antibiotics: classes and mechanism of action

Antibiotics are kind of antimicrobials commonly used to treat bacterial infections. They can either be produced naturally from other organism or by synthesis, and they act both killing or inhibiting the growth of bacteria⁴. They are commonly classified on the basis of their mechanism of action.

1.2.1. Antibiotics targeting the cell wall

Bacterial cells are surrounded by a cell wall made of a sugar polymer, the peptidoglycan, which results from the cross-linking of glycan strands by the action of a transglycosidases. The penicillin binding proteins (PBPs) allow the cross-linking between the D-alanyl-D-alanine portion of the peptide chains, which extend from the sugars in the polymers, and form cross-links with glycine residues⁵.

Drugs like beta-lactam antibiotics and glycopeptides act inhibiting the cell wall synthesis. Since beta-lactam ring mimics D-alanyl-D-alanine portion, it has been proposed that PBPs interacts with the antibiotic and it is not available for the synthesis of new peptidoglycan, leading to cell wall lysis. Also glycopeptides binds to D-alanyl D-alanine, preventing the binding of the subunit with the PBP.

1.2.2. Inhibition of protein synthesis

This class is heterogeneous and it includes different compounds able to inhibit protein biosynthesis by targeting the 30S or 50S subunit of the bacterial ribosome, including aminoglycosides, tetracyclines or macrolides. The aminoglycosides (AG) are positively-charged molecules which attach to the outer membrane leading to formation of pores allowing itself to penetrate inside the bacterium. AG's interact with the 16S r-RNA of the 30S subunit near the A site through hydrogen bonds, causing misreading and premature termination of translation of mRNA. Tetracyclines, such as tetracycline, chlortetracycline, doxycycline, or minocycline, act upon the conserved sequences of the 16S r-RNA of the 30S ribosomal subunit to prevent binding of t-RNA to the A site, while chloramphenicol interacts with the conserved sequences of the peptidyl transferase cavity of the 23S r-RNA of the 50S subunit, preventing binding of t-RNA to the A site of the ribosome. Compounds like macrolides, lincosamides, and streptogramins B affect the translocation, by targeting the conserved sequences of the peptidyl transferase center of the 23S r-RNA of the 50S ribosomal subunit. Oxazolidinones inhibit protein synthesis by binding to 23Sr RNA of the 50S subunit and suppress 70S inhibition and interact with peptidyl-t-RNA.

1.2.3. Folic acid metabolism inhibitors

There are two types of molecules interfering with the biosynthetic pathway of the folic acid metabolism: sulfonamides and trimethoprim.. The former inhibit the dihydropteroate synthase in a competitive manner with higher affinity for the enzyme than the natural substrate, p-amino benzoic acid, while the latter inhibits the enzyme dihydrofolate reductase.

1.2.4. Inhibitors of DNA replication

Quinolones are a group of broad spectrum synthetic antibacterial active against both Gram positive and Gram negative bacteria^{6,7}. They act by inhibiting the activity of two essential bacterial type II topoisomerases paralogues, DNA gyrase and topoisomerase IV, which are involved in the modulation of the chromosomal supercoiling required for DNA synthesis, transcription and cell division.

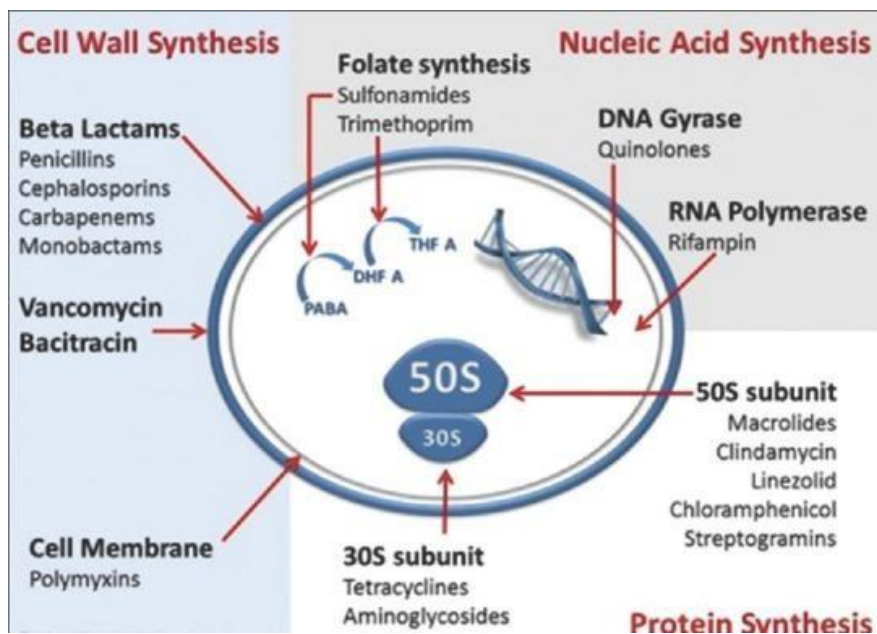


Fig 1.2. Class of antibiotics according to their mechanism of action⁵.

1.3. Quinolones and fluoroquinolones

The quinolones are a family of antibiotics containing a bicyclic core structure related to the compound 4-quinolone (figure 1.3) . Nowadays four generations of compounds have been developed.

Nalidixic acid is considered as the first quinolone antibiotic, followed in 1962 by a 6-fluoroquinolone synthesized by the Imperial Chemical Industries (ICI)⁸. Quinolones became a widely used drug class in the 1980s with the development of a second generation of compounds, the fluoroquinolones, in which the structure was modified with the introduction of a fluorine at the six position and a major ring substituent at position seven⁷ (figure

1.3), resulting in the increase of their activity. The second generation includes enoxacin, norfloxacin, and ciprofloxacin, which were further modified by the

addition of a piperazine ring to the R7 position and addition of a cyclopropyl group to the R1 position (figure 1.3). This combination made ciprofloxacin the first choice used against *Pseudomonas aeruginosa* today.

The third-generation quinolones are commonly used in the treatment of community-acquired pneumonia, acute sinusitis and acute exacerbations of chronic bronchitis. They include levofloxacin, gatifloxacin, moxifloxacin and sparfloxacin, which are active against gram-positive organisms, like *S. pneumoniae*, and atypical pathogens such as *Mycoplasma pneumoniae* and *Chlamydia pneumoniae*⁹. These drugs result in the addition of several functional groups to the R7 position (alkylated piperazine), R5 (-NH₂, -OH, and -CH₃) and R8 (Cl)⁸ (Figure 1.3).

The fourth generation class added significant antimicrobial activity against anaerobes⁹ due to the presence of nitrogen at the R8, while a 2,4 difluorophenyl group at the N position improves the overall potency of the drug⁸.

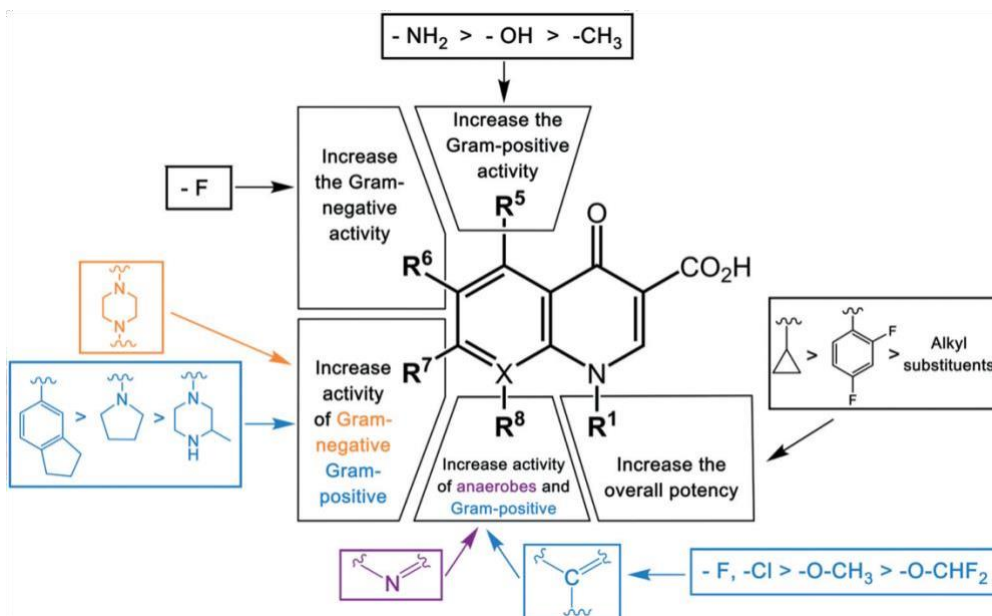


Fig.1.3. Structure-activity relationships of quinolones drugs⁸.

1.3.1. Quinolones target: DNA gyrase and topoisomerase IV

Quinolones act by inhibiting the activity of two essential type II topoisomerases, DNA gyrase and topoisomerase IV, which are involved in the modulation of the chromosomal supercoiling^{7,8}. DNA gyrase uses the energy of ATP hydrolysis to actively introduce negative supercoils into DNA, relieving the torsional stress that accumulates in front of replication forks and promoting local melting needed for transcript initiation by RNA polymerase. Topoisomerase IV is only able to relax positive supercoils and its major function is decatenation of the interlocked daughter chromosomes at the end of replication^{7,10,11}.

1.3.2. Structure of topoisomerase II enzymes

The general structure of type II topoisomerases is an A₂B₂ heterotetramer composed of two pairs of identical subunits, GyrA and GyrB for gyrase, ParC and ParE for topoisomerase IV^{7,10}.

The B subunit of *E. coli* DNA gyrase (GyrB) corresponds to the ParE subunit topoisomerase IV and to the N-terminal half of the eukaryotic enzymes, whereas the gyrase A subunit (GyrA) aligns with the ParC subunit and the C-terminal half of the eukaryotic enzymes¹².

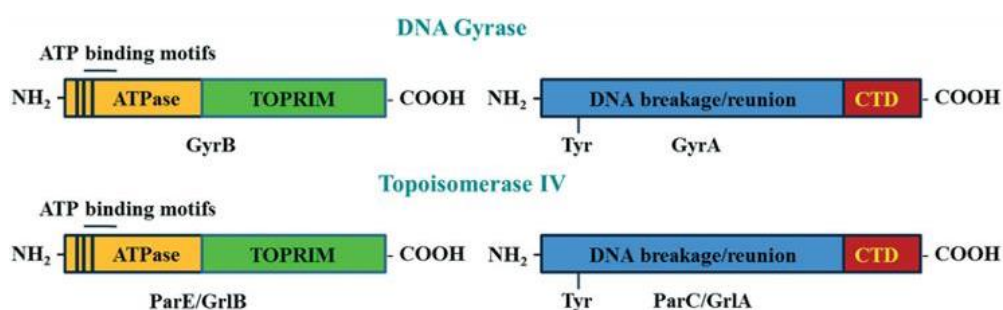


Fig 1.4. Sequence comparisons of topoisomerase IV and DNA gyrase⁸.

We can see that ParC and GyrA subunits consist of an N-terminal part, comprising a winged-helix domain (WHD), a tower domain, and a C-terminal beta-pinwheel domains (CTD) (figure 1.5, region coloured in green, blue, and purple).

These domains bind to DNA and promote DNA breakage^{10,13,14}.

The N-terminal and C-terminal regions of the highly conserved ParE (GyrB) subunits form the ATPase- and Mg²⁺-binding-TOPRIM domains, respectively (figure 1.5, yellow and red regions)¹⁰. TOPRIM domains bind with the WHDs region, forming the tetramer.

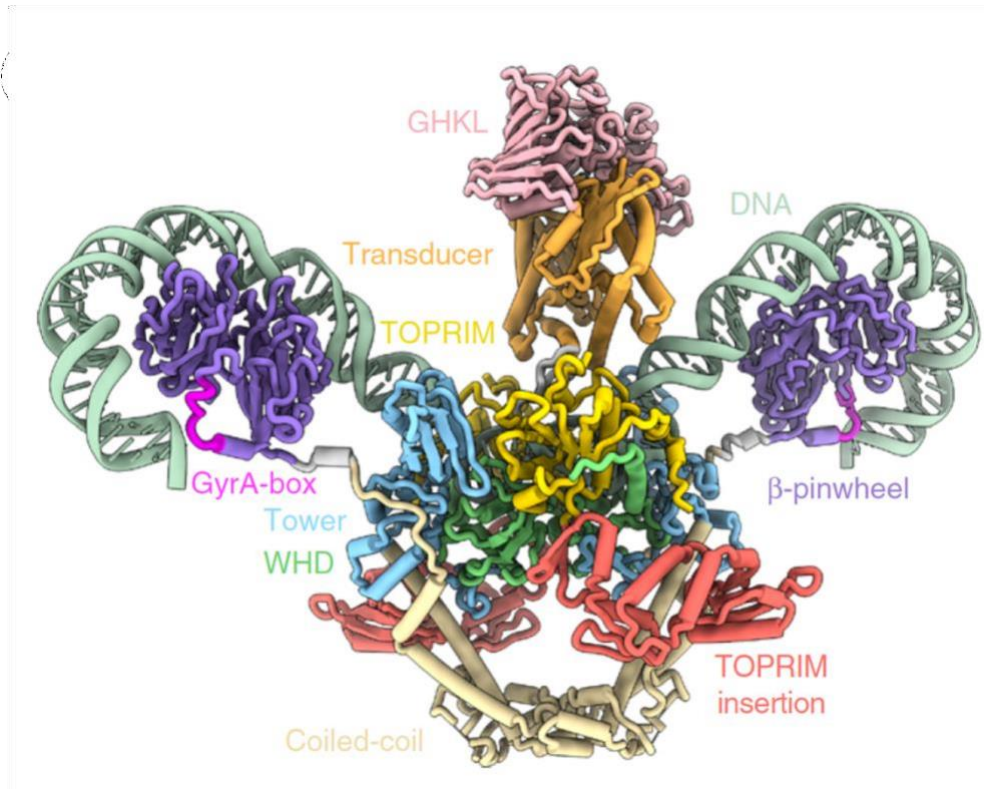


Fig 1.5. (A) Molecular structure of a bacterial topoisomerase II enzyme in complex with DNA (light green). Protein is colored according to the different domains¹⁵.

The main function of topoisomerases II is the formation of a transient DNA break involving a covalent-enzyme DNA intermediate named the ‘cleavage complex’¹⁰.

It has been proposed that the enzyme acts as a protein clamp that captures DNA as described in detail in Figure 1.6.

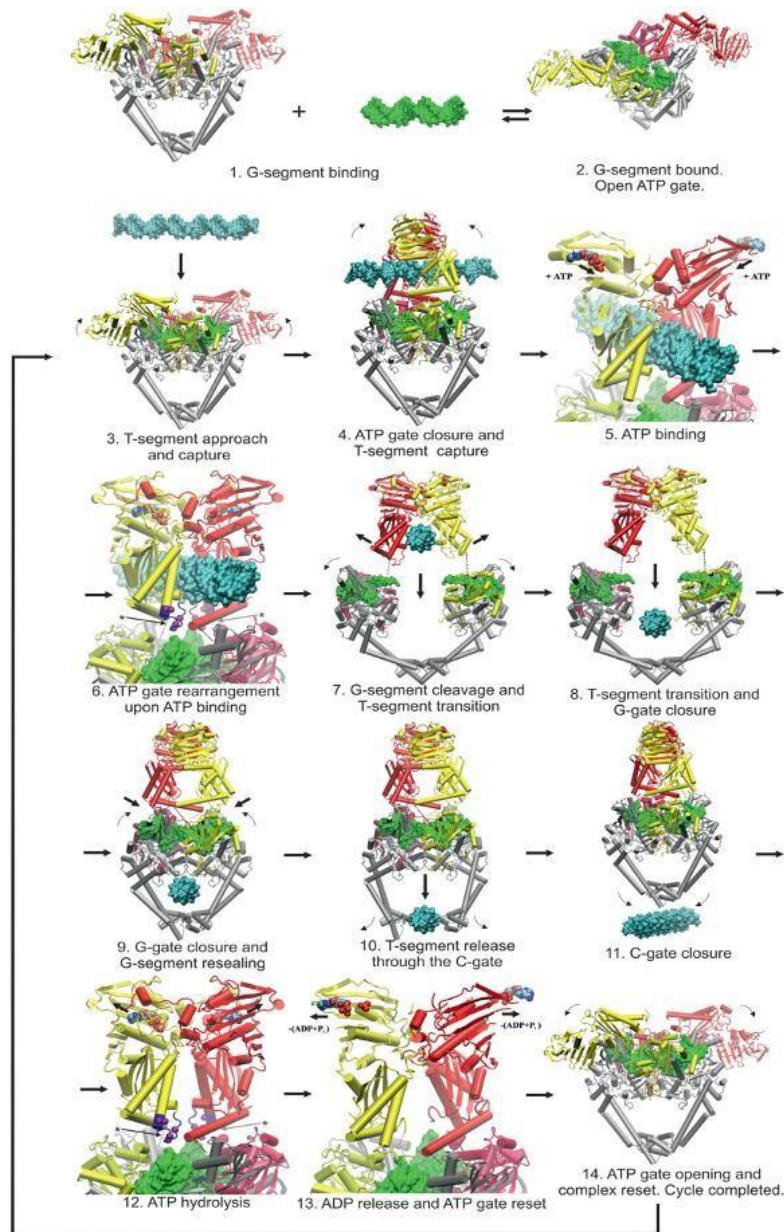


Fig.1.6. Proposed model for topoisomerase II mechanism of DNA capture and transport¹⁰. The protein, in its open

conformation (the 'open clamp') binds the G- DNA (1-2). Binding of the ATP leads to dimerization of ATPase domains resulting in the closing of the N-gate and capturing of the T-DNA(3-6). The DNA T-segment moves through the opened DNA gate, then the G-segment is resealed and the C-gate is opened, leading to the release of the T-Segment(7-10). The closure of the C-gate and release of ADP opens the N-gate, converting the closed conformation to the open conformation, ready for another cycle(11-14).

1.3.3. Mode of action of quinolones drugs

Quinolones interfere with the topoisomerase II mechanism by reversibly binding to the cleavage complexes at the enzyme–DNA interface in the cleavage–ligation active site^{7,8,16} in a non-covalent manner⁸, leading to the formation of a quinolone–topoisomerase–DNA ternary complex that causes the DNA replication machinery to become arrested at blocked replication forks, resulting in an inhibition of DNA synthesis. Moreover, when the DNA tracking systems collide with these drug–DNA cleavage complexes, permanent chromosome breaks are generated, triggering the DNA stress responses which activates DNA repair enzymes. The effect is bacteriostasis at low concentrations of quinolone or bactericidal activity at lethal concentrations^{7,8}.

Studies on the X-ray crystal structures of topoisomerase IV/gyrase- drug- DNA complexes from different organisms

revealed that quinolone is hemi- intercalated into each DNA strand and stacked against the DNA bases at the cleavage site (Figure 1.7). The drug molecule in close register to ParC S79/ GyrA S81 and ParC D83/GyrA E85 residues (*S. pneumoniae* numbering), and with a cluster of two glutamates and an arginine (E474, E475 and R456) on ParE/GyrB subunit^{16,17}. Moreover, the binding occurs through a water– metal ion bridge, where a noncatalytic Mg²⁺ ion coordinated with four water molecules forms a bridge for hydrogen bonding between the quinolone and the serine and acidic residues that act as anchor points to the enzyme^{7,16}.

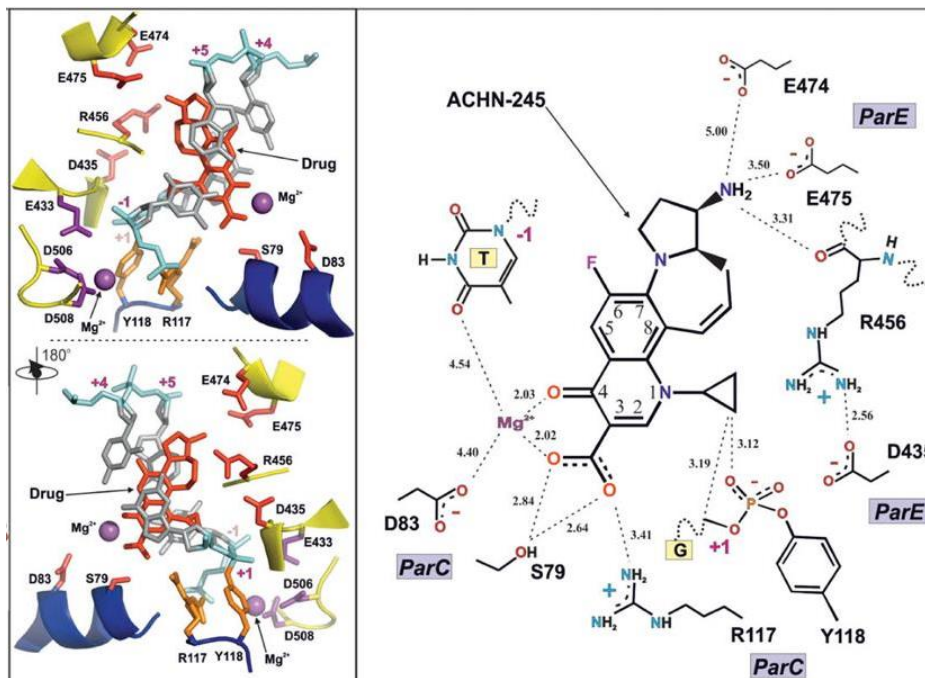


Fig 1.7. Details of binding mode of quinolone compound ACHN-245. Figure modified from the original work of Laponogov et al¹⁶.

1.4. The antibiotic resistance

Antibiotic resistance (AR) is a natural process whereby microorganisms acquire the ability to resist the effects of antimicrobial drugs, due to the selective pressure that results from exposure to these compounds^{18,19}. It is a global public health threat and it is estimated to cause around 300 million premature deaths by 2050 with a loss of up to \$100 trillion to the global economy²⁰.

There are two general strategies for resistance. One comprises mechanisms that transfer resistance vertically from a bacterium to its progeny, for example mutations in gene(s) often associated with the mechanism of action

of the compound, the other includes the acquisition of external genetic determinants of resistance, likely obtained from intrinsically resistant organisms present in the environment, through horizontal gene transfer like transformation, conjugation or transduction^{19,20}.

During thousands of years of evolution, bacteria have evolved sophisticated mechanisms of drug resistance to avoid killing by antimicrobial molecules²⁰. These mechanisms include:

- modification of the antibiotic molecule, for example by enzymes (acquired from horizontal gene transfer or synthesized by the bacterium itself) that perform chemical reactions like acetylation, phosphorylation, adenylation or the destruction of the drug: β -lactamases are able to destroy the amide bond of the β -lactam ring, rendering the antimicrobial ineffective;
- modification of the drug permeability. Bacteria have developed mechanisms to prevent the antibiotic from reaching its intracellular or periplasmic target by decreasing the uptake of the antimicrobial molecule. Permeability modification can be achieved with the expression of efflux pumps which actively extrude the quinolone or passively by increasing the expression of porins;
- target site alterations, for example bacteria can produce proteins like TetM which interacts with the ribosome and remove the tetracycline from its binding site in a GTP-dependent manner, or PBP2a, a PBP that has low affinity for all β -lactams, including penicillins, cephalosporins and carbapenems. But one of the most common alterations are point mutations of the protein target, like the ones occurring in the rpoB protein which confer resistance to rifamycin, or the well-known point mutations associated with quinolone resistance.

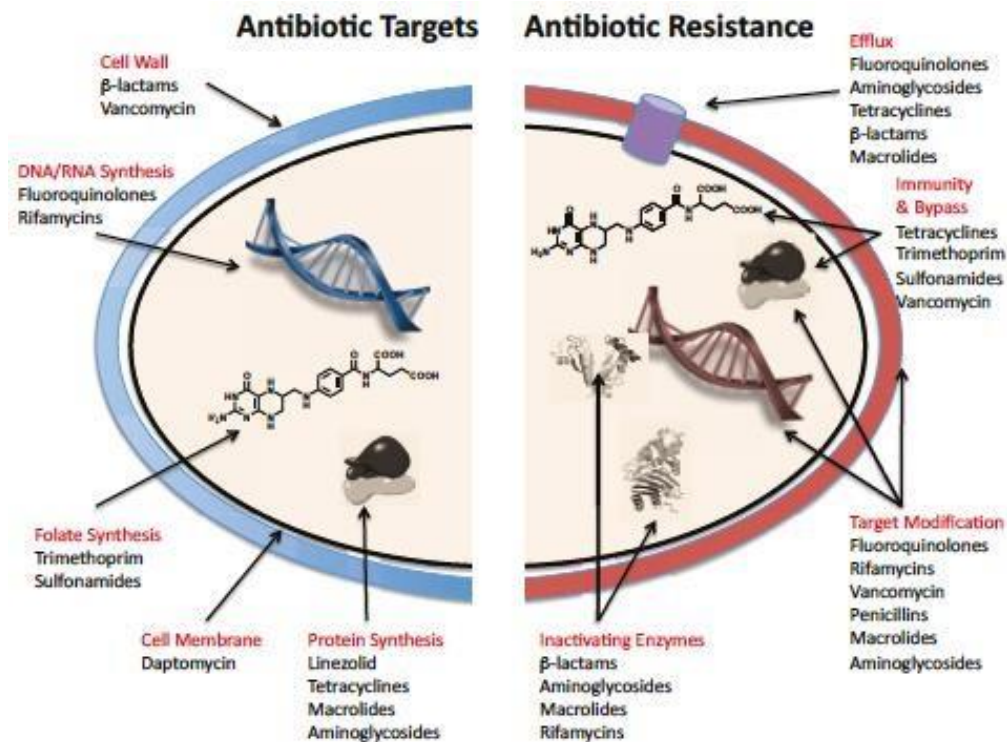


Fig 1.8. Common mechanisms of antibiotic resistance⁷.

1.5. Mechanisms of quinolone resistance

The acquisition of quinolone resistance is associated with: plasmid-acquired resistance genes synthesizing either proteins capable of protecting the drug target(s), or drug-modifying enzymes or drug efflux pumps pumps^{7,8}; other chromosomal mutations lead to reduced drug accumulation by either decreased uptake or increased efflux; chromosomal mutations that alter the quinolone target enzymes thus decreasing their affinity for the drug.

1.5.1. Plasmid-mediated quinolone resistance (PMQR)

This kind of resistance is due to plasmids carrying genes of resistance such as the Qnr, a protein of the pentapeptide repeats family competing with DNA in the binding to the topo IV and gyrase enzymes thus inhibiting the quinolone from entering the cleavage complex and reducing the number of double-stranded breaks on the chromosomes. This results in reduced quinolone toxicity to the chromosomes. Also a derivative of aminoglycoside acetyltransferase is able to acetylate the drug, thereby decreasing the quinolone activity^{7,8,21}. Finally QepA, a plasmid-mediated efflux pump, is able to decrease susceptibility to hydrophilic fluoroquinolones, especially ciprofloxacin and norfloxacin²¹.

1.5.2. Reduced drug accumulation

In Gram positive bacteria, quinolone resistance by increased efflux is due to the overexpression of three efflux pumps, NorA, NorB and NorC, causing four to eight fold increase in bacterial resistance to quinolones. In Gram negative organisms, the reduced or loss of expression of porins such as OmpF, OmpC, OmpD and OmpA or, the overexpression of OmpX, a downregulator of porin expression has been implied in increased antibiotic resistance to quinolones and other drugs²¹.

1.5.3. Mutations within the QRDRs

The most common mechanism of high-level resistance in different bacteria species is due to mutations occurring in a specific region of the topoisomerase IV/ gyrase subunits, termed the quinolone resistance

determining region (QRDR)²¹. Mutations in this region result in amino acid substitutions that structurally change the target protein and the drug-binding affinity of the enzyme⁷. The most common mutations are located on GyrA/ParC subunits and affect serine and acidic residues (aspartic or glutamic acid). Several bacterial species present similar mutations at equivalent positions⁸. In *S. pneumoniae* the replacement of a serine residue in position 79 of ParC with a phenylalanine or tyrosine introduces a bulky amino acid side chain, which interferes allosterically with drug binding¹¹, while in the substitution in alanine, the non-polar side chain prevents the interaction with the ligands and/or with the magnesium ion, with the subsequent loss of affinity²².

The additional mutation of the aspartic acid in position 83 of ParC can lead to the disruption of the metal ion water bridge required for the quinolone- protein interaction^{7,23}. Mutations of acidic residues in the QRDR of GyrB as well as in ParE in *E. coli* and other species have been shown to confer quinolone resistance, suggesting that they may interfere with charge interactions between drug and target²¹, for example in *E. coli* it has been proposed that the loss of negative charges in GyrB

subunit makes it difficult for a positively charged group of quinolones to associate with the protein²².

1.6. Bioinformatics resources to fight antimicrobial resistance

Given the increasing crisis represented by antimicrobial resistance (AMR), several bioinformatics tools and databases have been developed in order to better understand the underlying molecular mechanisms²⁴. Improvements in next-generation sequencing technologies and computational methods are facilitating rapid antimicrobial resistance gene identification and characterization in genomes and metagenomes²⁵.

Bioinformatics approaches can be categorized as those that focus on prediction of AMR, for example identifying the presence of resistance genes, and in those that study the mechanism of resistance, using gene expression profiles, metabolomics, structural analysis and so on²⁴.

Several methods are sequencing – based , in which assemblies, genomic contigs or full gene sequences are annotated for resistance determinants by comparing them against antimicrobial resistance reference databases, like the software PointFinder^{24,26}.

Other methods avoid genome assembly and directly map reads (or k-mers) to the reference databases using pairwise alignment tools such as Bowtie²⁵.

Also machine learning algorithms have been explored to predict the presence of genes of resistance. For example logistic regression was implemented to differentiate between vancomycin-susceptible and vancomycin-intermediate *Staphylococcus aureus*²⁷; Mykrobe predictor uses k-mer screening to identify resistant SNPs and genes in *M. tuberculosis* and *S. aureus*²⁸, while RAST uses AdaBoost classifier and PATRIC database for resistant genes annotation²⁹, and DeepARGs is a new tool which uses deep learning to identify resistant genes³⁰ in several different bacteria.

Name	Accessibility	Year	Description
<i>Assembly-based tools</i>			
Resfinder ³¹	Web and/or standalone	2012	Acquired AR genes identification
ARG-ANNOT ³²	Standalone	2014	AR genes identification
RG1 ³⁰	Web and/or standalone	2015	Predict resistomes, AR genes and point mutations
ARGs-OAP ³³	Web and/or standalone	2016	Pipeline for AR genes detection
ARIBA ³⁴	Standalone	2017	AR genes identification
PointFinder	Web and/or standalone	2018	AR point mutations identification
NCBI-AMRFinder ³⁵	Standalone	2018	AR genes and point mutations identification
<i>Read-based tools</i>			
SRST2 ³⁶	Standalone	2014	Virulence and AR genes identification
SEAR ³⁷	Web and/or standalone (archived)	2015	Pipeline for AR genes identification
ShortBRED ³⁸	Standalone	2015	Protein families profiling
PATRIC ²⁹	Web	2016	Genomic analysis of bacterial pathogens
SSTAR ³⁹	Standalone	2016	AR genes predictor
KmerResistance ⁴⁰	Web	2016	Gene identification
GROOT ⁴¹	Standalone	2018	AR genes profiling
DeepArgs ⁴²	Web	2018	AR genes identification with machine learning

Table 1.1. List of different tools developed for the antibiotic resistance identification. Table modified from the original published by Boolchandani et al²⁵.

Other important resources are public databases collecting information about known genetic determinants of resistance and

information from multiple studies that include antimicrobial susceptibility testing. Generalized databases (like CARD or ARDB) deal with mechanism information and cover several classes of antibiotics, while specialized databases focus on a specific compounds or bacterial species, for example Lactamase Engineering Database (LacED) provide information on β -lactamases, or MUBII-TB-DB which provides information on resistance in *Mycobacterium tuberculosis*²⁵.

1.6.1. The Comprehensive Antibiotic Resistance Database

CARD³⁰ is a curated database which provides nucleotide and protein sequences of genes of resistance, a resistant SNPs database, and a controlled vocabulary, the Antibiotic Resistance Ontology (ARO). The ARO is organized in three branches: determinant of Antibiotic Resistance (ARO:3000000), antibiotic molecule (ARO:1000003) and mechanism of antibiotic resistance (ARO:1000002)⁴³. Each new AMR determinant is manually curated by a dedicated team, and the process includes review of the scientific literature, and adding of annotation from external publications. As of September 2019 CARD included 4336 ontology terms, 2923 AMR determinants, 1304 resistant variant mutations and 2648 curated publications.

Moreover, CARD offers its own tool, known as Resistance Gene Identifier (RGI), which predicts AMR genes and mutations from submitted genomes using different tools such as Prodigal BLAST or DIAMOND and curated resistance mutations

included with the AMR detection model³⁰. RGI can detect functional homologues of antimicrobial resistance proteins and mutations conferring antimicrobial resistance²⁵.

1.6.2. PointFinder tool and database

PointFinder has been developed for the detection of point mutations associated with drug resistance and it is an extension of ResFinder, a well-known web server for the identification of acquired antimicrobial resistance genes.

PointFinder consists of a nucleotide database with reference sequences (nucleotide sequences of genes susceptible to antibiotics) and a point mutation database containing information on codon positions and substitutions. Given a query sequence, the tool uses BLASTn to match the sequence in the nucleotide database, then the program goes through each alignment comparing each position for the query (sequence found in input sequence) with the corresponding position in the subject (database sequence). All mismatches are compared with the point mutation database²⁶ and the variations related with AR are highlighted.

1.6.3. NCBI-AMRfinderPlus database

NCBI-AMRfinderPlus³⁵ is a tool for the identification of acquired resistance genes using NCBI's curated AR database and curated collection of Hidden Markov Models (HMMs).

This database derives from the Pathogen Detection Reference Gene Catalog, a non-redundant database of bacterial genes related to antimicrobial resistance. This includes highly curated, AMR-specific genes and proteins from the Bacterial Antimicrobial Resistance Reference Gene Database (BioProject PRJNA313047) and point mutations³⁵.

1.7. Structural analysis of point mutations

Amino acid substitution is one of the basic events that can drive evolution, leading to a variety of consequences on protein stability and function⁴⁴ or interfering with the binding of a drug to its target. Point mutations in a protein sequence may result in a change or loss of the native structure or the binding site, which in turn may cause a change or loss of function, and ultimately may yield different phenotypes⁴⁵. These effects depend on various factors, including the type of protein and the structural context in which it occurs. For these reasons, structural analysis is necessary to understand and ideally predict the effects of a mutation⁴⁶.

Amino acid substitutions can have locally repercussions but also long ranges effects. The replacement can introduce unfavorable hydrophilicity or hydrophobicity, or charges shift, or even modify the relative amino acid solvent accessibility⁴⁷. Moreover salt bridges and hydrogen bonds can be affected⁴⁸, like also electrostatic, charge–dipole and dipole–dipole interactions⁴⁶.

Eventually, mutations can cause conformational changes perturbing the energy landscape⁴⁹ and affect the binding affinity of the protein to the drug⁴⁶.

1.8. *Streptococcus pneumoniae* as case study

Also known as pneumococcus, this bacterium is a Gram positive, extracellular, opportunistic pathogen, regular common colonizer of the upper respiratory tract^{50,51}. It causes frequent infections associated with the airways, such as otitis media, sinusitis and bronchitis and it can be spread through airborne transmission⁵¹. *S. pneumoniae* is found predominantly in the mucus layer overlying the epithelial surface of the upper respiratory tract, pneumolysin induces inflammation which stimulates secretions and increases shedding and bacterial load. In this way pneumococcus is spread in the environment and it can colonize a new host by evading clearance mediated by IgA1(immunoglobulin on mucosal surface) via a pneumococcal zinc metalloprotease ZmpA. Then the bacterium expresses two enzymes, peptidoglycan-N-acetylglucosamine deacetylase (PgdA) and attenuator of drug resistance (Adr), that modify its peptidoglycan and render it resistant to the lytic effects of lysozyme facilitating colonization are adherence to host cells and tissues, and evasion of clearance by mucociliary flow. Local spread, aspiration or seeding to the bloodstream results in invasive inflammatory diseases⁵¹.

According to the World Health Organization (WHO), the bacterium is the fourth most frequent microbial cause of fatal infection, and the most common cause of bacterial pneumonia and

meningitis⁵⁰. Included as one of 12 priority pathogens, the continued high burden of disease and rising rates of resistance to penicillin and other antibiotics have renewed interest in prevention of the pneumococcal infections⁵¹.

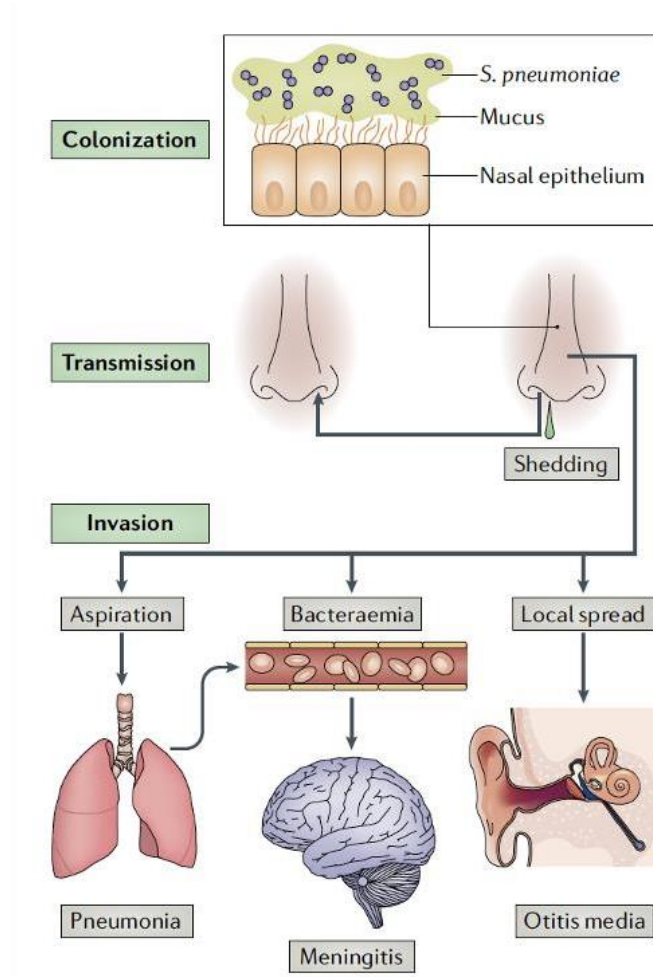


Fig 1.9. The life cycle of *Streptococcus pneumoniae* and the pathogenesis of pneumococcal disease⁵¹.

1.9. Machine learning

Machine learning (ML) is a branch of artificial intelligence that is able to learn from experience in order to predict future events or scenarios that are unknown to the computer⁵².

Experience exists in the form of training datasets, which the machine learner uses to build a general mathematical model about that domain.

Nowadays, ML is used in a large number of bioinformatics areas such as genomics, proteomics, microarrays, systems biology, evolution, text mining^{53,54}.

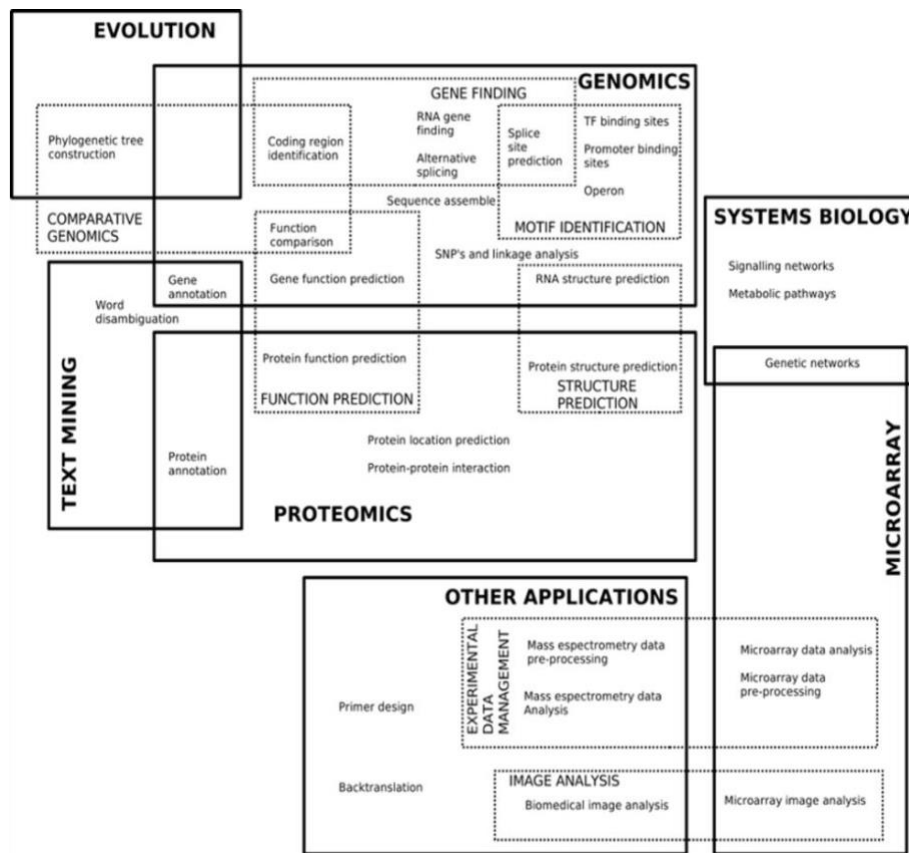


Fig.1.10. Application fields of machine learning in bioinformatics⁵⁴.

ML algorithms may be broadly classified as **Supervised learning**, a learning mechanism that infers the underlying relationship between training data and a target variable, minimizing the error for a given set of inputs. The training data comprise feature vectors and a desired output value (the class label).

Unsupervised learning algorithms are designed to discover hidden structures in unlabeled datasets, in which the desired output is unknown. The goal of ML in this case is to hypothesize representations of the input data for efficient decision making, forecasting, and information filtering and clustering⁵².

There are several steps in the development of a machine learning algorithm:

- 1) Data collection
- 2) Preprocessing of data, like formatting, cleaning by removing missing data or by normalization/standardization, sampling to remove redundancy
- 3) Transformation of the data specific to the algorithm, for example feature scaling or decomposition
- 4) Training on the dataset and evaluation on the test set to verify its effectiveness and performance
- 5) Application of the validated model to perform an actual task of prediction

1.9.1. Common ML algorithms employed in bioinformatics

Several algorithms are widely used in bioinformatics such as logistic regression, support vector machines, classification trees, random forest, and nearest neighbour⁵⁴. Also deep learning is being incorporated in vast majority of analysis pipelines, due to the advent of the big data era in biology⁵⁵.

For example clustering algorithms have been successfully applied to gene expression analysis on sequence data from tumors⁵⁶. Random Forest has gained popularity and it is becoming a common standard tool, especially in the context of low-dimensional data⁵⁷. Databases that store DNA, RNA, protein sequences and macromolecular structures are growing exponentially. The size and complexity of these data require the use of advanced computational tools, like neural networks⁵⁸.

1.9.2. K-means clustering

Known as Lloyd's algorithm, the K-means clustering clusters data by separating a set of N samples into K disjoint subsets (or clusters) S_j , each described by the mean μ_j . The means are called 'centroids' of the cluster.

$$J = \sum_{j=1}^K \sum_{n \in S_j} |x_n - \mu_j|^2$$

The algorithm consists of a two-step re-estimation process: first data point are assigned to the cluster whose centroid is closest to that point, then each centroid is recalculated to the mean of all points assigned to it. These two steps are repeated until a threshold is reached, such that there is no further change in the assignment of data points. In other words, it is repeated until the centroids do not move significantly^{52,59}. K-means clustering was used in the structural analysis of point

mutations related to antibiotic resistance, in order to identify clusters of mutations with structural commonalities.

1.9.3. Random forest

Random forest (RF) is a class of methods that use a classification tree as the base classifier⁵⁴. As the name suggests, it consists of a number of individual trees that works as an ensemble: n predictors are combined to solve a classification or estimation problem through averaging⁵². The basic unit (base or weak learner) is a binary tree constructed using recursive partitioning scheme and it is typically grown using the methodology of the Classification and Regression Tree: starting from the root node, the process involves splitting among all the possible splits at each node. The resulting child nodes are the purest^{52,60}. RF uses a two-stage randomization for the growth. Instead of using all the variables to split a tree node, the algorithm selects at each node of each tree a random set of variables that are used as candidates to find the best split⁶⁰.

The RF can be summarized in a set of steps⁵²:

- 1) From the original dataset n bootstrap samples are selected to construct B trees.
- 2) For each sample a tree is grown.
- 3) At each node of the tree:
 - a) A subset of features is randomly selected

- b) Features that provide the best split perform the binary split on that node
- c) The next node selects another set of variables and performs the preceding steps
- 4) Take the majority vote of all the B subtrees.

In simple terms, each individual tree makes a class prediction: the class with the most votes becomes the model's prediction. Compared with logistic regression, k-means and neural networks, random forest was the best performing machine learning algorithm on our dataset of structural features of point mutations associated with antibiotic resistance.

1.9.4. Artificial neural networks

Artificial neural networks (ANNs) are algorithms vaguely inspired by biological neural networks⁶¹. The basic unit is the artificial neuron or '*perceptron*', a classifier that, using a threshold activation function, separates two classes by a linear discrimination function⁵⁴.

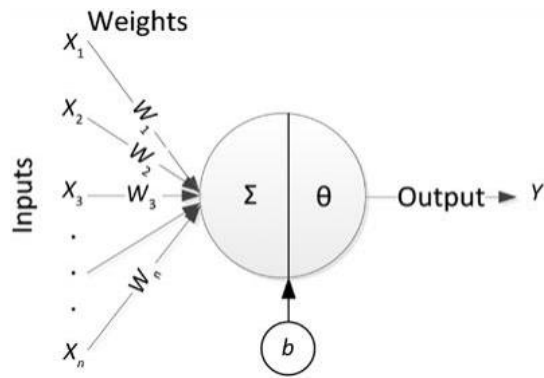


Fig 1.11. Figure of an artificial neuron⁵².

The inputs are connected to the neuron through weighted connections emulating the dendrite's structure. The summation (Σ), the bias (b), and the activation function (θ) play the role of the cell body. The propagation of the output is analogous to the axon in a biological neuron⁵².

In the ANN, the perceptrons are connected together in consecutive layers. A layer of neurons is a "column" of neurons that operate in parallel, the output of the layer is the vector output, which is formed by the individual outputs of neurons.

In order to perform classification, the ANN is trained with a non-linear function which express the hidden relationship between the features (x) and the label (v), training the parameters (w) and making the model fit the data⁵⁵. The standard algorithm adopted is the forward-backward propagation^{52,54,55}. At the beginning the network is initialized randomly, then the network run and the output is compared with the target value to get the difference (loss or error). Then, the error back-propagation and optimization

help adjust the parameters of the model, making the output as close to the target value as possible.

The most commonly used activation functions are ReLU for the inner layer and Softmax for the output layer, loss functions are cross-entropy for classification and mean squared error for regression. The optimizers include stochastic gradient descent (SGD), Momentum, Adam⁵⁵.

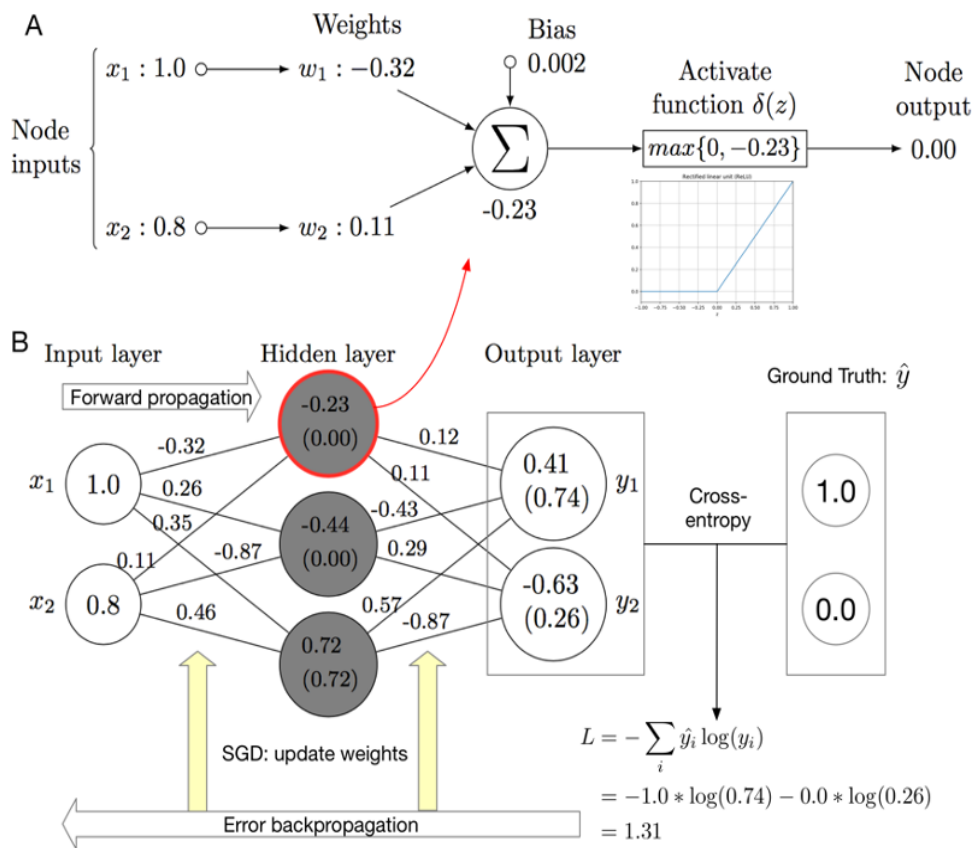


Fig 1.12. A) Operations inside a single node, B) explanation of the forward- backward propagation in a ANN composed of an

input layer, an hidden layer of three neurons and an output layer⁵⁵.

We tested ANN on our dataset but random forest proved to be more accurate.

1.9.5. Evaluation methods

A possibility when developing predictive models is that the algorithm is over-fitted on the existing data, resulting in a drastic performance drop when it is applied in practical studies with novel data⁶². An example evaluation method is cross-validation, in which all data are used as both training and test dataset. It is considered as a compromise solution when the number of available samples is very limited⁶².

In binary predictors there are two classes (0 or 1, True or False, resistant vs susceptible ...) and for each sample in the test set we have a real label and a predicted label. The real label indicates the class the sample really belongs to, while the predicted label is the output of the predictor⁶². We can count the outcomes in the form of false positive (FP), true positive (TN), false negative (FN) and true negative (TN) and represent the combinations of predicted and actual values in a table called **confusion matrix**^{52,62}.

		Prediction	
		Positives	Negatives
Real	Positives	<i>TP</i>	<i>FN</i>
	Negatives	<i>FP</i>	<i>TN</i>

Fig 1.13. Confusion matrix, positive and negatives are the two possible classes, TP= outcome is correctly identified as positive, TN= outcome is correctly identified as negative, FP=outcome is incorrectly identified as positive, FN=outcome is incorrectly identified as negative.

2. Aim of the thesis

The extensive use (and misuse) of antibiotics has led to the spread of resistant bacterial pathogen strains, causing a severe problem worldwide. Action must be taken quickly in order to stem this situation.

Quinolones are a new class of antibiotics that bind bacterial topoisomerases and inhibit bacterial cell replication. They have been important in limiting the spread of penicillin- and macrolides-resistant bacteria like *Streptococcus pneumoniae*. However, alarmingly, resistance to quinolones has recently appeared in *S. pneumoniae* strains and other bacteria.

According to the World Health Organization (WHO), *S. pneumoniae* is the fourth most frequent microbial cause of fatal infection, and the most common cause of bacterial pneumonia and meningitis. Due to the threats this bacterium poses especially for the elderly and the children, *S. pneumoniae* was chosen as a case study for this work.

This PhD project focuses on characterizing the molecular mechanisms of the resistance to quinolones caused by the appearance of point mutations, and in using this information to develop bioinformatics tools for both the analysis and the inference of point mutations associated with antibiotic resistance (AR). In this thesis, for brevity, we will refer to point mutations associated with AR as “resistant mutations” and to residue types

appearing in reference or susceptible sequences as “wild-type residues” or “susceptible residues”.

Specifically, the aim of this thesis consists in:

1. performing a sequence and structural analysis of the mutations involved in the resistance to quinolones in *Streptococcus pneumoniae* and in other bacterial species. Since quinolones bind to the gyrase and topoisomerase enzymes involved in DNA replication, these enzymes were studied in detail, their three-dimensional structure was modelled where needed and “resistant mutations” were mapped onto the structure;
2. developing and testing machine learning methods for the detection and prediction of mutations involved in antibiotic resistance;
3. developing and deploying a database collecting structural information on point mutations associated with resistance to quinolones;
4. developing, testing and deploying a web server for the structural analysis, characterization and visualization of variants detected in sequence positions associated with quinolone resistance.

3. Materials and methods

3.1. Topoisomerase and gyrase sequences

Reference nucleotide sequences for the non-pathogenic *S. pneumoniae* R6 strain were retrieved from the NCBI nucleotide database⁶³ with the following accession IDs:

NC_003098.1:752241-754721 for the **parc** gene,

NC_003098.1:749887-751830 for **pare**,

NC_003098.1:c1097931-1095463 for **gyra**,

NC_003098.1:715818-717764 for **gyrb**.

Reference protein sequences for *S. pneumoniae* R6 strain were retrieved from the NCBI protein database⁶³ with the following accession IDs: NP_358351.1 for **ParC**, NP_358350.1 for **ParE**, NP_358692.1 for **GyrA**, NP_358309.1 for **GyrB**. Protein sequences associated with quinolone resistance of other bacteria (see table 4.1) were obtained from CARD ‘model variants’ database³⁰ and redundant sequences were removed. Susceptible sequences were obtained from UniprotKB-Swiss-Prot using the proteins name (‘GyrA’, ‘ParC’, ‘GyrB’, ‘ParE’) as keywords and selecting only proteins from bacteria. Records with the keyword ‘antibiotic resistance’ and ‘quinolone’ were added to the CARD resistant sequences. Among the susceptible sequences, we chose the topoisomerase IV and gyrase subunits of the organisms analysed as reference sequences from which we extracted the susceptible amino acids. When it was not possible to retrieve the

wild-type reference sequence for a given bacterium from UniprotKB-SwissProt, the reference susceptible sequence deposited in CARD was used. We collected all the sequences with the point mutations causative of drug resistance from CARD, but we noticed that for some variations, the corresponding resistant protein sequence with the mutation was not present in the database. For this reason, we mutagenized in silico the susceptible reference sequence from UniprotKB-SwissProt, replacing the wild type amino acid with the resistant variant.

3.2. Pairwise and multiple sequence alignment

The Blast suite (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>)⁶⁴ was used to perform several tasks. **MakeblastDB** was utilized to generate the Blast reference nucleotide and protein databases for the topoisomerase IV and gyrase proteins. **Blastx** was implemented in the Quinores 3D web server to translate the nucleotide query sequence into the corresponding amino acid sequence using the Bacterial, Archaeal and Plant Plastid genetic code. **Blastp** also was incorporated into the Quinores3D web server and in Quinores3D_pred to perform pairwise alignment between the protein query sequence and the reference sequence. Multiple sequence alignments of resistant (MSAs) and susceptible sequences were carried out using Muscle 3.8.31 tool⁶⁵ and Clustal Omega web server (<https://www.ebi.ac.uk/Tools/msa/clustalo/>)⁶⁶. Muscle

software was also installed in the web server to run MSA using the user's input sequence and the susceptible reference on the fly.

For the interpretation of the alignments, we used the Jalview version 2 software, a free program for multiple sequence alignment editing, visualization and analysis⁶⁷. MSAs were represented with the Clustal color scheme, which highlights amino acids according to their properties (e.g. hydrophobic residues are shown in blue, positively charged in red, polar in green). Also, from Jalview we annotated the amino acid property conservation (measurement of the conservation of physicochemical properties in a MSA column) and the alignment quality.

Since Jalview cannot be embedded in an HTML page, in order to show pairwise alignments and MSAs generated by Quinores3D, we incorporated a Javascript sequence alignment viewer, JSAV⁶⁸, in the HTML output page. The sequence numbering of the reference is shown above the alignment, conserved residues are represented as dots, and mutations using 1-letter code (Figure 4.7 A).

3.3. Homology modelling protocol

No experimentally determined structures comprising the whole protein are available neither for the *S. pneumoniae* topoisomerase IV nor for the gyrase. For this reason, four

separate homology models were built for the subunits ParC, ParE, GyrA and GyrB. Homology modelling was performed with the software MODELLER v9.23⁶⁹ and templates' searching was carried out with the HHpred⁷⁰ server.

Protein sequences were downloaded from the NCBI with IDs NP_358351.1 for ParC, NP_358350.1 for ParE, NP_358692.1 for GyrA, NP_358309.1 for GyrB.

These sequences were submitted to the HHpred server to search for three – dimensional protein structures in Protein Data Bank (PDB)⁷¹ to be used as templates. Yet, full homologous protein structures are not present in PDB from other bacteria. We chose to generate a multi-template model for each subunit, including a template covering the N-terminal and one the C- terminal regions. From the HHpred output, templates were chosen taking into consideration first the structures with lowest Evalue, and then among them the structures with the best resolution.

For GyrA the experimentally determined 3D structure PDB 4Z2C (chain A, resolution 3.19 Å) and 1SUU⁷² (resolution 1.75 Å) were chosen as templates for the N-terminal and C- terminal, respectively; for GyrB, 4Z2C (chain C) and 3ZKB⁷³ (resolution 2.90 Å, chain I); for the ParC subunit of topoisomerase IV PDB 3RAE⁷⁴ (chain B, resolution 2.90 Å), PDB 1ZVU (resolution 3 Å) and 1WP5⁷⁵ (resolution 1.79) were chosen as a templates for the N-terminal and C- terminal part

respectively, while for ParE 3RAE chain C and 5J5P⁷⁶ (resolution 1.97 Å, chain B).

To run MODELLER and evaluate the results we used the version implemented in Chimera, setting the number of models as 20 and selecting also the ‘thorough optimization’ option, which optimizes more thoroughly than the default one.

Resulting models were chosen based on the Discrete Optimized Protein Energy (DOPE) score⁷⁷ and on the QMEAN Z-score⁷⁸ calculated with the SWISS-MODEL web server⁷⁹. Loops were refined with the ‘DOPE method’ implemented in MODELLER. We also built the full protein – DNA – magnesium – quinolone complexes for topoisomerase IV and gyrase with Chimera by importing DNA, quinolone and magnesium coordinates from the 3D structure of 3RAE for the former and the structure of 4Z2C for the latter.

Protein	Sequence ID	template ID	Chain	Resolution (Å)
ParC	NP_358351.1	3RAE	B	2.90
		1WP5	A	1.79
		1ZVU	A	3.00
ParE	NP_358350.1	3RAE	C	2.90
		5J5P	I	1.97
GyrA	NP_358692.1	4Z2C	A	3.19
		1SUU	A	1.75
GyrB	NP_358309.1	4Z2C	C	3.19
		3ZKB	A	2.90

Table 3.1. List of PDB IDs, chains, and related resolution of the proteins chosen as templates for the four subunits. Protein sequence IDs are from NCBI.

Homology modelling was also implemented in Quinore3D and in machine learning pipeline in order to generate the 3D structures starting from resistant and susceptible sequences, but the protocol was modified: instead of a multi-template model, we used the previous generated protein models as templates; moreover, we generated just one model using the ‘fast/approximate’ option. These modifications were required in order to speed up the process of model building and simplify its integration inside the machine learning and Quinore3D finder pipeline.

3.4. Structural analysis

3.4.1. Quinolone binding site analysis

The interactions between residue side chains and the drug were analyzed with Chimera both for the wild type and the mutated proteins. Variations and the residues at 5 Å of distance were selected with the selection (*‘sele’*) command. Inside a selection, hydrogen bonds are calculated with the *findhbond* function, whereas contacts and clashes are detected using the *findclash* function with the following parameters: overlap -0.4 and allowance 0.0 for contacts, overlap 0.6 and allowance 0.4 for clashes.

Findclash identifies interatomic clashes and contacts based on Van der Waals (VDW) radii, where clashes are unfavorable interactions where atoms are too close together and contacts are all kinds of direct interactions such as polar and nonpolar, favorable and unfavorable (including clashes). The overlap between two atoms (i,j) is defined as the sum of their VDW radii (VDW_r) minus the distance between them (d) and minus an allowance, that takes into account the contribution of probable hydrogen-bonded pairs:

$$overlap_{ij} = VDWr_i + VDWr_j - d_{ij} - allowance_{ij}$$

3.4.2. Electrostatic analysis

Electrostatic analysis was carried out in a region of the protein, comprising the position of interest and all residues within 5 Å from it. Ligands were not taken in consideration and therefore removed.

Charges and potentials were calculated with the APBS (Adaptive Poisson- Boltzmann Solver)⁸⁰ software, which requires as input a PQR file, that is a modified PDB file with charges calculated for each atom. This file was generated with the PDB2PQR software using the PARSE force field and the PDB file containing the residues at 5 Å from the mutation as input.

From the PQR file, APBS generated the potential map and calculated the charge density and electrostatic potential.

3.4.3. Relative solvent accessible surface analysis

The relative accessibility surface (RASA) for each single amino acid in the protein model was calculated using the function ‘SASA’ in Biopython, which runs the DSSP⁴⁷ software needed for this computation.

3.4.4. Protein structure visualization and analysis

For the interactive visualization and analysis of molecular structures, Chimera software was used⁸¹. This tool was widely utilised to show the topoisomerase IV and gyrase structures, build the 3D models, map the variations, analyze the interactions between quinolones and the side chains of the mutated amino acids, show bonds and contacts between atoms. Chimera was also used to generate high quality figures shown in Results.

Due to the fact that Chimera GUI cannot be used directly in a web page we implemented the Quinores3D code with the WebGL (<https://www.khronos.org/webgl/>) applications for molecular visualization. These applications were embedded as libraries in the result web page to display the protein of interest, map the mutations on the structure, show distances from the residues and drug or magnesium, show the quinolone binding site and the interactions (H-bonds, contacts), and display the electrostatic surface.

3.5. Primer design

Quinores3D Primers allows users to design forward and reverse PCR primers given a nucleotide sequence using the open-source Primer3 software⁸².

Results are outputted in a table reporting, for each primer, the sequence, the content in GC, melting temperature, stability of any basepairing of that primer to itself, stability of any basepairing of the 3' end of the primer to itself, formation of hairpin loops.

We generated a set of primers (forward and reverse) specific for mutagenesis purposes. The nucleotide sequences carrying the quinolone resistance variations were submitted to Primer X web server

(<https://www.bioinformatics.org/primerx/documentation.html>), a tool developed for the automated design of mutagenic primers for site-directed mutagenesis. Primers with optimum values of length and GC content were collected into the Quinores3D Database, 'primers' table.

3.6. Annotation of mutations and information retrieval

Several sources were used to gather information and annotate mutations, including literature searching and use of specialised databases.

Information was retrieved from the literature by searching Pubmed with terms like 'Streptococcus pneumoniae',

'quinolone', 'resistance', 'mutations', 'not efflux'. Scientific articles were manually downloaded and inspected; mutations occurring in the quinolones targets were manually annotated and information about the Minimal inhibitory concentration (MIC) and the effect of amino acid replacements on drug-binding were collected.

CARD, PointFinder and NCBI pathogens databases were also downloaded and quinolone resistance variations reported for bacteria other than *S. pneumoniae* were extracted. All the raw data were organized in a dataframe using the Pandas library. Then after careful inspection, data were re-organized as the Quinores3D Database.

Quinores3D Database is a relational database developed in MySQL version

8.0.21. The database is organized in six tables (refer to Table 3.2):

Table	Description
eitable	Values from electrostatic analysis
homologtable	Homologous position and known resistant mutations in bacteria
intable	Mechanisms of resistance
mictable	Values of MIC calculated for different strain
primermutable	Site specific mutagenesis primers

sasatable	Values for relative accessible surface
-----------	--

Table 3.2. Short description of Quinores3D database tables.

3.7. Web server technical specification

The Quinores3D server has been deployed on the Cloud@ReCaS-Bari (<http://cloud.recas.ba.infn.it/>) IaaS (Infrastructure as a Service) cloud platform, on a virtual server with 4 virtual CPUs, 8 GB of RAM and a public IP address. Three virtual volumes (120 GB total disk space) have been added to the virtual machine, to store data of the MySQL database, of the BLAST database, and of the Apache web server, respectively.

The choice of deploying the service on cloud resources has been made in order to warrant scalability and elasticity, since it is possible to transparently and quickly enlarge or shrink the resources assigned to the virtual machine, and split the components of the logical architecture (web server, DBMS, scripts) on different servers, if needed. This approach will also allow in the future the possibility to easily migrate the service to another cloud resource provider, without any lock-in problem. Future plans include the refactorization of the involved software using a microservice architecture, in which each component is executed in a Docker⁸³

container on a Mesos/Chronos cluster, and the web frontend is deployed on a different machine with respect to the one(s) hosting the MySQL and BLAST databases, which are isolated on private networks.

The creation of the corresponding Docker images is currently undergoing. All code is open source and fully available on GitHub platform (<https://github.com/>) that provides hosting for software development and versions control.

3.8. Data analysis and code scripting

Python version 3.6.5 (<https://www.python.org/>) was used for data manipulation, parsing, analysis and web server development. The Python CGI (Common Gateway Interface) module was used to write the main code of the web server, to generate dynamically webpages and to interact with external programs incorporated in the server. Pandas library (<https://pandas.pydata.org/>) was used for data analysis while Matplotlib (<https://matplotlib.org/>) for data representation and graph plotting.

For sequence and structure analysis, Biopython was used⁸⁴. It was largely implemented in the web server but also for all the script developed for the mutations analysis or for the machine learning preparation, using modules such as SeqIO for sequence parsing, AlignIO for Blast result and MSA manipulation, PDB module for the structural analysis.

3.9. Application of information theory on sequence analysis

Conservation analysis from a MSA can be used for predicting functionally important residues in protein sequences and in ligand binding⁸⁵. Also variability provides important information about proteins⁸⁶.

Shannon entropy (H) can be used as a measure of residue diversity and residue conservation⁸⁷. It can be defined as a measure of uncertainty about the identity of objects in an ensemble:

$$H(X) = - \sum_{i=1}^N P_i \log P_i$$

where P_i was the probability of given amino acids and N was the number of letters in a sequence⁸⁸. To normalize between 0 and 1 the logarithm is taken to base 20.

In our work, Shannon entropy was applied to study the conservation of the key positions involved in drug resistance among the bacteria.

3.10. Machine learning

Machine learning algorithms were developed with the Python library Keras (<https://keras.io/>) and scikit-learn⁵⁹ using Google Colab⁸⁹ service.

3.10.1. Feature encoding for the predictor model

Each amino acid was represented as a vector of numeric descriptors. We used the AAindex database⁹⁰ and the iFeature package⁹¹ to encode the following features: hydrophobicity, hydrophilicity, side chain mass, residue volume (BIGC670101), steric parameter (CHAM820101), SASA in folded structure (CHOC760102), molecular weight (FASG760101), size of side chain (DAWD720101), normalized Van der Waals volume (FAUJ880103), Net charge of amino acid (KLEP840101), hydrophathy index according to Kyte – Doolittle (KYTJ820101), Polarity (GRAR740102), Bulkiness (ZIMJ680102), Side-chain contribution to protein stability (kJ/mol) (TAKK010101), Free energy of solution in water, kcal/mole (CHAM820102), charge transfer capability (CHAM830107_list). Moreover, relative solvent accessibility calculated with DSSP, electrostatic potential, net charge, charge density calculated with APBS at a range of 5 Å, distance from drug and magnesium were added as structural features. Finally, we divided the amino acids according to their physicochemical properties into 5 groups, assigning for each group a unique number (apolar:0, aromatic:1, polar:2, positively charged:3, negatively charged:4). This number was added as additional descriptor.

3.10.2. K-means clustering

Cluster analysis was performed with the K-means algorithm implemented in scikit-learn with the following parameters: max iterations 300, relative tolerance $1e^{-4}$, initial step 10. The number of clusters was chosen using the ‘elbow method’⁸³ implemented in the kneed library ([‘https://pypi.org/project/knead/’](https://pypi.org/project/knead/)). For clustering the following features were selected: relative solvent accessibility, electrostatic potential, net charge, charge at a range of 5 Å, distance from drug and magnesium, residue volume (BIGC670101), hydrophobicity index according to Argos et al., 1982 (ARGP82010), free energy of solution in water, kcal/mole (CHAM820102), steric parameter (CHAM820101), residue accessible surface area in tripeptide (CHOC760101), relative mutability (DAYM780201), solvation free energy (EISD860101), molecular weight (FASG760101), polarity (GRAR740102), side chain volume (KRIW790103), hydropathy index according to Kyte – Doolittle (KYTJ820101), surrounding hydrophobicity in folded form (PONP800101), amphiphilicity index (MITS020101), bulkiness (ZIMJ680102), radius of gyration of side chain (LEVM760105), and amino acid groups as descriptors.

3.10.3. Training, validation and test set

For both clustering and predictor models the data set was represented by the residues collected as described above, for a total of 362 residues. Among them, 22 were removed because their phenotypic effect was unknown. After the structural analysis

we decided to also remove positions that were unlikely to be involved in drug resistance, for a final set of 201 resistant variations and 115 wild type ones.

For the classifier algorithms, we oversampled the original data with the SMOTE function from Imbalanced learn package (<https://imbalanced-learn.readthedocs.io/en/stable/>), in order to obtain a proportion of 50% resistant and 50% susceptible residues.

The dataset was randomly splitted into training and validation sets, with a proportion of 70% for the training and 30 % for the validation set, using the *sklearn train_test_split* function. The test set was obtained taking all the resistant and wild type sequences from CARD and UniprotKB SwissProt that were not used in the training set.

Point mutations were selected as described above. In order to avoid incorrect assessment, we randomly choose 500 resistant and 500 susceptible residues for the test set.

Data were normalized using the *MinMaxScaler function* from *sklearn*.

3.10.4. Random forest and neural network

model The random forest classifier was set with the following parameters: `n_estimators=300`, `criterion='entropy'`, `n_jobs=-1`, `random_state=0`.

The neural network model is composed of three hidden dense layer:

- First dense layer with shape 16 and 'ReLU' activation function
- Dropout layer with a dropout rate of 20%
- Second dense layer of shape 8 and 'ReLU' activation function
- Output layer with 1 node and sigmoid activation function

For the compiler the 'adam' optimizer was used and for the loss measure the 'binary crossentropy' function. Accuracy and loss for training and validation were used as metrics to evaluate the performance of the model.

Model was trained for 200 epochs with a batch size of 32.

3.10.5. Metrics and statistical tests

For the structural analysis we adopted the Mann–Whitney U test implemented in the SciPy library⁹² to select the most relevant features. Features with a p-value higher than 0.05 were considered not statistically significant and they were not included in the analysis.

For the evaluation of our machine learning models, k -fold cross validation, accuracy, logarithmic loss, F1-score, precision, recall, confusion matrix and ROC curves were used.

In the k-fold cross-validation, the dataset was randomly partitioned into k parts of equal size. Each fold is left out of the design process and used as a test set. The model is trained k times,

for each k-round the k-th part is used as the test dataset, while the remaining k – 1 parts form the training dataset.

After all k rounds of training and testing, every sample in the dataset was used as a testing sample once and only once. The prediction performance can be estimated by averaging the prediction results over the whole dataset^{52,54,62}. We adopted a 5-fold cross validation, in which our training and test set were evaluated 5 times.

Referring to the Introduction section for the definition of false positive (FP), true positive (TN), false negative (FN) and true negative (TN), the performance measures can be defined as :

- *Accuracy (AC)*

The number of data points correctly classified by the classification algorithm.

$$AC = \frac{TP + TN}{TP + TN + FN + FP}$$

- *Precision (P)*

The frequency of true positives among all positive outputs.

$$P = \frac{TP}{TP + FP}$$

- *Recall (R)*

Recall or sensitivity is the frequency of correctly predicted positive samples among all real positive samples.

$$R = \frac{TP}{TP + FN}$$

- *F1 Score*

F-measure measure Recall and Precision at the same time using the Harmonic Mean.

$$F1 = \frac{2 \cdot R \cdot P}{R + P}$$

It can be used to compare two classifiers.

ROC curve describes the relationship between the sensitivity and false positive rate (FPR). The FPR can be defined as

$$FPR = \frac{FP}{FP + TN}$$

Given a scoring scheme, the values of R and FPR will change along with the threshold values and for every cut-off value, a dot can be plotted (coordinates FPR,R). The curve connecting the dots is the ROC curve, with a diagonal that is called the line of no-discrimination. The more the ROC curve is close to the

diagonal, the more the predictor is close to a random guess, whereas the more it is close to the top left corner, the more the performance is good⁶². We can use the area under the curve (AUC) to measure the performance of the predictor. AUC of an ROC curve is equal to the probability that a randomly selected positive sample gets higher scores than a randomly selected negative sample.

4. Results

4.1. Three-dimensional modelling of topoisomerase IV and gyrase protein structures

Full three-dimensional (3D) structures of the *Streptococcus pneumoniae* topoisomerase IV and gyrase in complex with DNA, magnesium and the drug (quinolone) are not available in the PDB database. Indeed, the available structures of the topoisomerase IV ParC and gyrase GyrA subunits lack the C-terminal regions, which comprises ~300 residues. The topoisomerase IV ParE and gyrase GyrB subunits lack the N-terminal region (~350 residues). In order to map all the mutations, a complete protein structure was required. 3D protein models were obtained using MODELLER⁶⁹, a tool developed for homology modelling. C-terminal and N-terminal protein templates were searched with the HHPRED web server⁷⁰, using the reference sequences for *S. pneumoniae* gyrases and topoisomerases IV proteins.

We generated the 3D models of the ParC, ParE, GyrA, GyrB subunits, each in complex with DNA, magnesium and the drug. For each sequence position associated with antibiotic resistance in *S. pneumoniae* (see Table 4.1) we also built models replacing the wild-type residue (corresponding to the non-resistant phenotype) with the other 19 possible amino acids. The same procedure (with little modification, see Materials & Methods)

was repeated for topoisomerase IV and gyrase of the bacteria showing antibiotic resistance to quinolones, generating a total of 413 models.

4.1.1. Topoisomerase IV ParC subunit model: a case study for the homology modelling procedure

The ParC subunit was chosen as case study to establish the protocol for the homology modelling.

The subunit is composed of two domains: an N-terminal domain of about 440 residues (from position 30 to 470 on the *S.pneumoniae* sequence), which contains the catalytic residues of the enzyme (see Introduction) and a C-terminal domain (residues from 500 to 800), which contains residues that are involved in non-specific DNA binding. Structures deposited in the PDB database only comprise the N-terminal domain in complex with DNA and the drug, so it is impossible to map and study a mutation if it falls in the C-terminal domain. As specified in Materials & Methods, we searched for templates by HHpred, selecting the PDB ID 3RAE chain B as a template for the N-terminal, and the PDB ID 1ZVU and 1WP5 as templates for the C-terminal region (see Table 3.1). These structures were used to generate 20 models, each of them was inspected with Chimera and submitted to the Swiss – model web server⁷⁹ to assess the quality of the model with the QMEAN function⁷⁸.

The best model has a QMEAN value of -3.56 (Figure 4.1.) It can be observed that the N-terminal has a good quality level (blue

region from residue 30 to 470), while local quality starts to decrease in the C-terminal region (residues colored in red, starting from residue 500). Positions related to drug resistance (79 and 83) are characterized by a good quality estimation.



Figure 4.1 Snapshot of the quality evaluation result from the Swiss – model web server⁷⁹. Protein sequence is colored according to local quality, from red (poor quality) to blue (high quality).

Finally, we reconstructed the protein in complex with DNA, magnesium and levofloxacin (a quinolone compound) by superposing our model to the protein 3RAE and importing the coordinates for the heteroatoms.

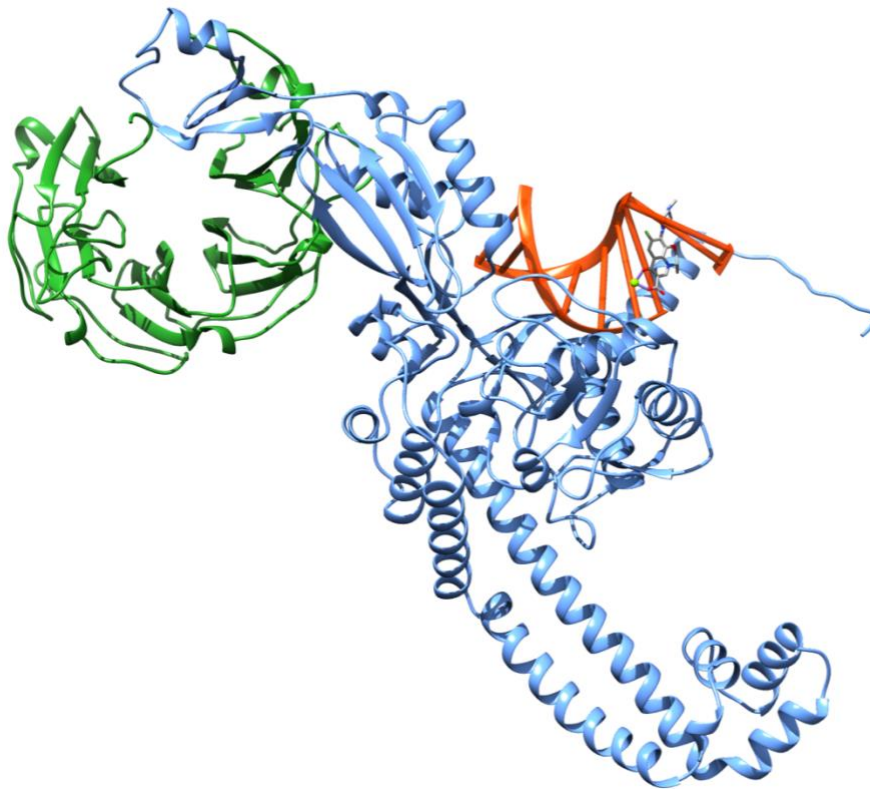


Fig 4.2 ParC model in complex with DNA (in orange), magnesium (represented as green sphere) and levofloxacin (represented as sticks). N- terminal domain is coloured in blue and the C- terminal in green.

4.2. Point mutation characterization and structural analysis

Amino acid substitutions may result in structural changes in the target protein, possibly affecting the drug-binding affinity of the enzyme. For this reason, we performed different kinds of structural analysis in order to evaluate the effects of the side chain replacements in protein positions associated with

resistance, with the aim of understanding if and how the modifications may potentially affect the interaction of the enzymes with the quinolone molecule.

In order to perform a structural analysis, the mutations were first mapped onto the protein structure (see Methods section 3.4.4), then the following properties were calculated: average distance between the residue side chain atoms and the drug atoms and the magnesium, types of bonds and interactions between the quinolone and the residue side chains, relative accessible surface area (SASA), electrostatic properties, and physico-chemical features (e.g. hydrophobicity).

4.3. Quinores3D: a web server for the structural analysis of the molecular mechanisms of resistance to quinolones

Quinores3D is a web server developed for the identification and characterization of point mutations associated with quinolone resistance in *S. pneumoniae* topoisomerase and gyrase proteins. The web server is hosted at the INFN ReCaS DataCenter and can be reached using the following URL

‘<http://bioinfoibpm.cloud.ba.infn.it/quinores3d/index.html>’.

It comprises a database (Quinores3D db), a tool to identify and analyse variations in topoisomerases/gyrases subunits (Quinores3D finder) and a tool to generate PCR primers specific for a gene of interest (Quinores3D primers).

4.3.1. Quinores3D db

Quinores3D db is a relational database specifically developed for the collection of data permitting the study of the molecular mechanisms of resistance to quinolones in *Streptococcus pneumoniae*. Quinores3D db contains a large amount of sequence and structural information on the point mutations occurring in GyrA, GyrB, ParC, and ParE sequence positions associated with bacterial resistance to quinolones. Table 4.1 reports the mutations and their corresponding sequence positions collected in Quinores3D db. The information was partly retrieved from the literature and partly obtained as the result of computational analyses. For each position related to quinolone resistance in the four gene (GyrA, GyrB, ParC, and ParE) sequences, the wild-type residue (i.e. the residue present in the reference sequence of the non-resistant strain) was artificially replaced with the 19 amino acids, and a 3D model of both the wild type and each mutated sequence was built, obtaining a total of 80 3D predicted structures. Structural annotation was generated for the wild type and each mutation (see Methods and Table 4.1).

Thirty seven mutations are known from both the literature and specialized resources (e.g. CARD³⁰) to be explicitly associated with quinolone resistance in *S. pneumoniae*. For each of these mutations, we also collected the values of the MIC from the bacterial strains carrying the mutation from available scientific studies. Moreover, we identified the type of interaction (e.g., hydrogen bond, water-ion bridge, electrostatic) established

between the residue side chain and the drug and the magnesium for both the wild type and the mutated residue, in order to study the structural changes introduced by the mutation. We also identified positions and variations associated with quinolone resistance homologous to the ones in *S. pneumoniae* from other resistant bacteria, resulting in a collection of 116 mutations from 18 organisms.

Bacteria	Sp. seq pos	Org. seq pos	Mutation
<i>Streptococcus pneumoniae</i>	GyrA		
	81	81	S --> C,I,F,Y
	85	85	E --> G,K
	GyrB		
	435	435	D --> N
	474	474	E --> K
	475	475	E --> K
	ParC		
	63	63	A --> T
	79	79	S --> A,I,L,F,Y
	83	83	D -->A,N,H,,V
	ParE		
	435	435	D --> N,H
	474	474	E --> K
475	475	E --> A	
<i>Acinetobacter baumannii</i>	GyrA		
	79	71	G --> C
	81	81	S --> L
	ParC		
83	79	S --> L	
<i>Bartonella bacilliformis</i>	GyrA		
	81	91	S --> A
	85	95	D --> N
<i>Burkholderia_dolosa</i>	GyrA		
	79	76	G --> D
	81	83	T --> I
	85	87	D --> H

<i>Campylobacter jejuni</i>	GyrA		
	81	86	T --> A,I,K,V
	85	90	D --> A,N,T,Y
<i>Capnocytophaga gingivalis</i>	GyrA		
	79	80	G --> N
<i>Clostridium difficile</i>	GyrA		
	81	82	T --> I,V
	GyrB		
	435	426	D --> N,V
	475	466	E --> K,V
<i>Enterococcus faecalis</i>	GyrA		
	81	83	S --> R,I,N,L,Y
	85	87	E --> G,K,L
	ParC		
	79	80	S --> R,I
	84	83	E --> A,K,T
<i>Enterococcus faecium</i>	GyrA		
	81	83	S --> R,N,I,L,Y
	85	87	E --> G,K,L
	ParC		
	79	80	S --> R,I
	84	83	E --> A,K,T
<i>Escherichia coli</i>	GyrA		
	81	83	S --> A,I,L,F,W,V
	85	87	D --> A,N,G,H,Y,V
	GyrB		
	435	426	D --> N
	ParC		
	79	80	S --> I,L,F,V,W
	83	84	E --> A,G,K,V
<i>Haemophilus parainfluenzae</i>	GyrA		
	81	84	S --> Y
	ParC		
	79	84	S --> F
	GyrA		
	81	83	S --> L,F

<i>Klebsiella</i>	ParC		
	79	80	S --> I
	83	84	E --> K
<i>Mycobacterium leprae</i>	GyrB		
	435	464	D --> N
	475	504	E --> V
<i>Mycobacterium tuberculosis</i>	GyrA		
	81	91	S --> A
	85	94	D -->
	GyrB		
	475	540	E --> V
<i>Mycoplasma genitalium</i>	GyrA		
	81	95	M --> I
	ParC		
	78	82	D --> N
	79	83	S --> I,R
	83	87	D --> G,N,H,Y
<i>Mycoplasma hominis</i>	ParC		
	79	91	S --> I

<i>Neisseria gonorrhoeae</i>	GyrA		
	81	91	S --> F
	85	95	D --> N,G
	ParC		
	79	87	S --> R
	83	91	E --> Q,G,K
<i>Propionibacterium acnes</i>	GyrA		
	81	101	S --> L
	85	105	D --> G
<i>Pseudomonas aeruginosa</i>	GyrA		
	81	83	T --> I
	85	87	D --> N,G,H
	ParC		
	79	80	S --> L

<i>Salmonella</i>	GyrA		
	81	83	S --> A,F,Y
	85	87	D --> N,G,K,Y
	GyrB		
	435	426	D --> N
	ParC		
	79	80	S -->R,I
83	84	E --> G,K	
<i>Staphylococcus aureus</i>	GyrA		
	81	84	S --> L
	85	88	E --> A,K
	ParC		
	79	80	S --> F,Y
	83	84	E --> G
	GyrB		
435	434	D --> N,H	
<i>Ureaplasma urealyticum</i>	ParC		
	79	83	S --> L
	83	87	E --> Q

Table 4.1. Variants and their corresponding sequence positions associated with drug resistance in bacteria collected in Quinores3D db. For each organism we reported the sequence position of the mutation(s) in the bacterium ('Org seq pos'), the homologous sequence position in *S.pneumoniae* ('S.p. seq pos') and the mutations in the format WT →list of amino acids causative of resistance.

Quinores3D db contains mutations associated with drug resistance in *S. pneumoniae*; for each variation, information is also collected about MIC, electrostatic potential and charges calculated in a range of 5 Å, as well as the relative solvent accessible surface, the mechanism of resistance, the homologous mutations in other

bacteria and the PCR primers specifically designed to amplify the nucleotide region containing the mutation.

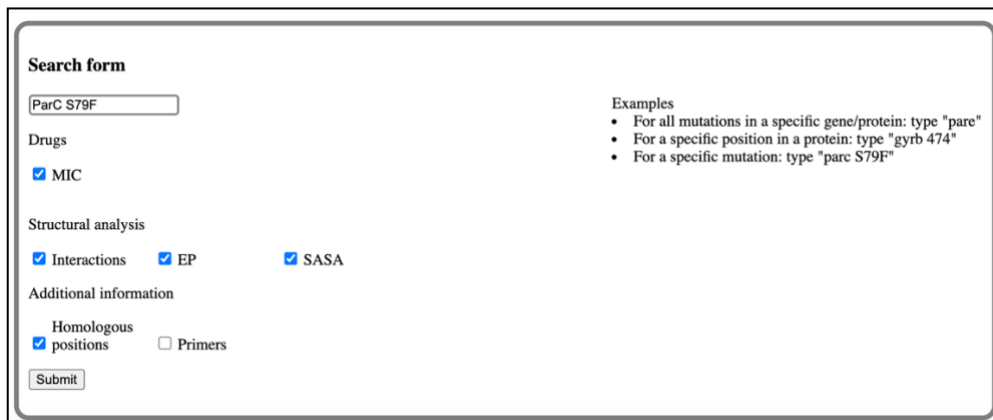
The database can be queried via a dedicated web page using keywords like the protein name, the position or a specific mutation. For example a user who is interested in the study of the mutation S79F in the topoisomerase IV ParC subunit , and wishes to retrieve the following information:

- the minimal inhibitory concentration (MIC)
- description of the molecular mechanism of resistance
- results of the structural analysis.

can type 'parc S79F' in the searching bar and mark the checkboxes 'MIC', 'Interactions', 'EP', 'SASA', 'Homologous positions' (Figure 4.4) .

Results can also be downloaded in tabular format as text files.

The result is a web page divided in different sections ('MIC', 'Structural information' ...) with the information organized in tables (Figure 4.4).



Search form

ParC S79F

Drugs

MIC

Structural analysis

Interactions EP SASA

Additional information

Homologous
 positions Primers

Submit

Examples

- For all mutations in a specific gene/protein: type "pare"
- For a specific position in a protein: type "gyrb 474"
- For a specific mutation: type "parc S79F"

Figure 4.3. Query example 'S79F' and 'ParC' as keywords. Users can mark the checkboxes to include MIC values, structural analysis (comprising interactions

and electrostatic analysis, and relative ASA calculation) and homologous positions in other bacteria.

Results found for search term "parc s79f"								
Drugs and MIC information								
MIC (minimal inhibitory concentration) values and quinolones information retrieved from literature								
Id	Position	Wt	Mut	Drug	MIC	PMID	Note	
parc_79_F	79	ser	phe	Ciprofloxacin	8	15317743	#none	
parc_79_F	79	ser	phe	Gatifloxacin	1	15317743	#none	
Click here to download MIC table.								
Structural information								
Structural information retrieved both from literature and computational analysis:								
<ul style="list-style-type: none"> • Interactions: direct effects of wild type amino acid or point mutation on drug binding • EP: value of the Electrostatic Potential calculated at 5 Å from the mutation. It can affect indirectly drug/magnesium binding. • Net_charge: value of the Net charge calculated at 5 Å from the mutation. It can affect indirectly drug/magnesium binding. • Charge_map: value of the Charge distribution calculated at 5 Å from the mutation. It can affect indirectly drug/magnesium binding. • SASA: value of the Relative Solvent-Accessible Surface Area calculated for the amino acid. It can affect indirectly drug binding. 								
Id	Position	Wt	Mut	Interactions	EP	Net_charge	Charge_map	SASA
parc_79_F	79	ser	phe	H_bond loss; steric hindrance	3493.0	-2.0	-25100.0	0,873096447
Click here to download Interactions table.								

Figure 4.4 Results for query 'ParC S79F' & MIC values, structural analysis, and homologous positions. 'Id': unique identifier for the mutation, 'Position' is relative to *S. pneumoniae* sequence numbering, 'Wt': wild type amino acid, 'Mut': mutation, 'Drug': compound for which the MIC ('MIC' field) is given. 'PMID': article identifier from which we retrieved the minimal inhibitory concentration, 'Note': annotation from the authors of the database, 'Interactions': the mechanism of resistance, 'EP', 'Net_charge', 'Charge_map': charge analysis, 'SASA': relative ASA respectively.

4.3.2. Quinores3D finder

Quinores3D finder can be used to identify amino acidic variations occurring in the topoisomerases and gyrases subunits.

The server home page is shown in Figure 4.3

Quinores3D finder

[QR3D Finder](#) [QR3D Database](#) [QR3D Primers](#) [About resistance](#)
[Documentation](#)

Description

Quinores3D Finder is a web server developed for the identification of mutations related to quinolone resistance in *Streptococcus pneumoniae*. It works comparing the quinolone protein target sequences (Gyrase:gyrA/gyrB, Topoisomerase:parC/parE) against the relative susceptible S.p. R6 strain protein sequences.
All the mutations found are mapped on the protein structure and compared with a quinolone resistance database.

INPUT FORM

Choose Quinolone Resistant Protein:
gyrA ▼

Select your type of input:
genome ▼

Upload your file: [Scogli file](#) Nessun file selezionato

OPTIONAL INFORMATION

Project name
email
Project description

START ANALYSIS

Examples

Click the button to download some input file examples
[Examples](#)

Figure 4.5. Quinores3D finder homepage.

It is structured in three sections: input form, optional information and examples. In the input form, users have to select the subunit

of interest (GyrA, GyrB, ParC, or ParE) and the type of input text file they will upload: a protein or nucleotide sequence, the complete bacterial genome or a list of point mutations. Table 4.2 reports the different input options and formats.

Input	Options	Description
Choose Quinolone Resistant Protein	GyrA, GyrB, ParC, ParE	The gyrase or topoisomerase subunit on which the user wants to perform the analysis
Select your type of input	genome gene protein mutation	DNA or protein sequence in fasta format (.fa, .fasta, .fna); a complete genome in fasta format (.fa, .fasta, .fna, .gbf); mutations described as AApositionAA (e.g. S70F) in text format (.txt)

Table 4.2. Quinores3D Finder input form options and formats.

Upon submission, the input file is pre-processed: nucleotide or genome sequences are firstly converted into protein sequences using Blastx. In the case of a genome sequence, the nucleotide sequence of the protein of interest is extracted from the genome with Blastn and then translated into protein. If a list of mutations is provided, the reference protein sequence available in Quinores3D db is mutagenized *in silico*, i.e. user-supplied variations are inserted in the reference sequence of the specified gyrase or topoisomerase subunit.

INPUT FORM

Choose Quinolone Resistant Protein:

Select your type of input:

Upload your file: Nessun file selezionato

Or paste your input:

```
L.Y.L.M.S.N.I.Q.N.M.S.L.E.D.I.M.G.E.R.F.G.R.Y.S.K.Y.I.I.Q.D.R.A.L.P.D.I.R.D.G.L.K.F.V.Q.R.R.I.L.Y.S
M.N.K.D.S.N.T.F.D.K.S.Y.R.S.A.K.S.V.G.N.I.M.G.N.F.H.P.H.G.D.F.S.I.Y.D.A.M.V.R.M.S.Q.N.W.K.H.R.E.I
I.V.E.H.I.G.N.G.S.H.S.D.P.P.A.A.M.R.V.T.E.A.R.L.S.E.I.A.C.V.L.L.Q.D.I.E.K.F.V.P.F.A.W.N.F.D.D
T.E.K.E.P.T.V.L.P.A.A.F.P.M.L.L.V.N.G.S.T.C.I.S.A.G.Y.A.T.D.I.P.P.H.N.L.A.E.V.I.D.A.A.V.M.I.D.H.P
T.A.K.I.D.K.L.M.E.F.L.P.G.P.D.F.P.T.G.A.I.I.Q.G.R.D.E.I.K.K.A.V.E.T.G.R.V.V.V.R.S.K.T.E.I.E.K.L.
```

Examples

Click the button to download some input file examples

OPTIONAL INFORMATION

Fig 4.6. Example of a web server run: the user uploaded the ParC protein sequence for the analysis.

A three-dimensional structure is generated by homology modelling from the user sequence with MODELLER⁶⁹. The identification of point mutations in the user sequence is carried out according to the following protocol: the user sequence is aligned with the reference using Blastp, then the resulting pairwise alignment between the user sequence and the reference sequence is parsed in order to identify the point mutations. All the variations found are compared with the mutations listed in Quinores3D db: if the mutation is known to be related to drug resistance, the corresponding information is retrieved from the database and displayed in the result web page, otherwise a structural analysis is carried out on the fly and the results are displayed in the result web page.

The output is a HTML page reporting the sequence alignment and the blast scores (score, sequence identity, sequence similarity, e-value, ...), a table with the mutations annotated and a viewer developed in WebGL showing the protein structure with the mutations identified in the query sequence. For each mutation occurring in the protein, a specific web page is generated, containing the results of the structural analysis. If the variation is known to be associated with quinolone resistance, the corresponding information is retrieved from Quinores3D db. The result page will display a customized description of the effect of the mutation in the form of a table reporting: the minimum inhibitory concentration (MIC) values retrieved from literature and specific for the bacteria containing the mutated protein, the

values of the residue relative accessibility to the solvent, values of the residue net charge, electrostatic potential and charge, as well as hydrogen bonds and contacts. Moreover, the user can explore the protein structure in great detail thanks to the molecular viewer embedded in the web page.

If the amino acid change is not described in the literature, but it occurs at a sequence position known to be related to drug resistance, a structural *ad hoc* analysis is performed on the fly.

The results page will be different: the MIC table is replaced by a summary table reporting structural values (solvent accessibility, net charge, etc..) for the wild-type and the mutated residue.

This allows users to see whether the mutation has introduced important structural changes or not. Moreover, distances from drug, magnesium and the QRDR positions are calculated and shown in the viewer.

Results

Blast result:

Bitscore	Evalue	Identity	Positives	Gaps
1671	0	99%	818/823	0/823

Sequence alignment:

Input sequence(_query) is aligned to the reference sequence (_ref) of the non resistant strain R6; the _query sequence line shows changed aminoacids (with respect to _ref) using their 1-letter code, unchanged-aminoacids as a ., ambiguos as "X" (if the codon translated could be an aminoacid or a stop codon)

Sequence numbering	10	20	30	40	50	60	70	80	90
	MSNIQNM S LE D IMGERFGRYSKYI I QDRALPDIRDGLKPVQRRILYSMNKDSNTF D KS Y RKSAKSVGNIMGNFHPHG D SS I YDAMVRMSQ N								
parc_ref									
parc_query	MSNIQNM S LE D IMGERFGRYSKYI I QDRALPDIRDGLKPVQRRILYSMNKDSNTF D KS Y RKSAKSVGNIMGNFHPHG D SS I YDAMVRMSQ N								

[Export Fasta](#)

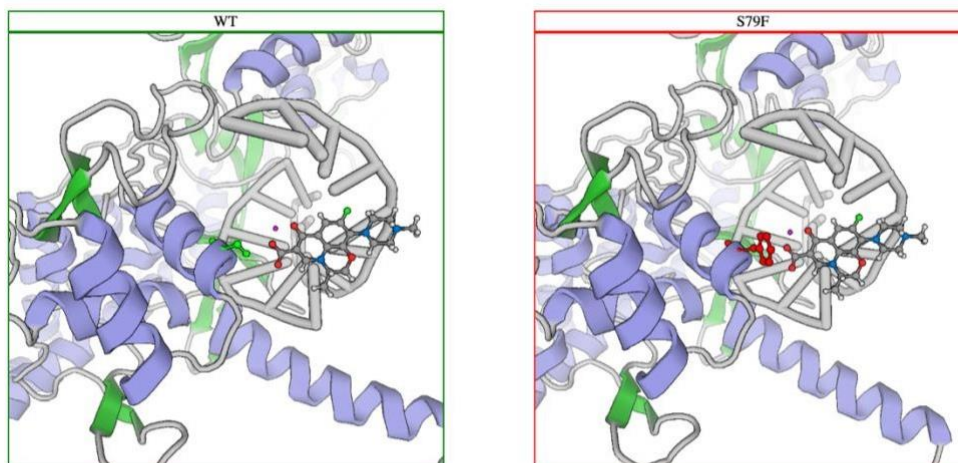
Mutations found: 5

Profile: 'resistance-associated; no resistance-associated; unknown'

Mutation (position according to reference sequence)	Profile	Note	Analysis Results
S79F	Resistant	mutation known for drug resistance	analysis
K137N	Unknown	No information from literature	analysis
R373H	Unknown	No information from literature	analysis
N473K	Unknown	No information from literature	analysis
A589E	Unknown	No information from literature	analysis

Fig 4.7.A) Output page for the analysis of a ParC resistant sequence. The sequence was carrying the mutation S79F known to be related with quinolone resistance. The output summarizes the blast result with the pairwise alignment and all the variations identified by comparing the sequence with the reference.

Structural analysis



Summary table

	Interactions	Solvent Accessibility Surface (Å ²)	Electrostatic energy(kJ/mol)	Net Charge (e)	Charge Distribution (e/Å ³)
Wylid type	hydrogen bonds(79S-OH --> drug-COOH)	0.63	5510	-2	-22800
S79F	H_bond loss; steric hindrance	0.87	3493	-2	-25100

Fig 4.7.B) Web server results: focus on the structural analysis. Left: wild type ('wt' green) and Right: mutant ('S79F' red). The table summarizes the structural analysis, e.g, mechanism of resistance ('Interactions'), charge analysis ('Electrostatic Energy', 'Net Charge', Charge Distribution') and relative solvent accessibility ('Solvent Accessibility Surface').

4.3.3. Quinores3D primers.

This tool can be used to generate PCR primers specific for the gene of interest using the open-source Primer3 software⁸² with default options. The output is a table with the sequence of the forward and reverse primers and other useful information such as temperature of melting, length of the primer, content of GC.

4.4. Analysis of point mutations associated with quinolone resistance

Due to the importance of point mutations in the rising of antibiotic resistance, several studies have been conducted including analyses of genomic sequences and/or structures, in order to better understand this complex phenomenon and face the challenge of developing new drugs.

Yet, most studies focus on single mutations in specific organisms, and so far a general study or classification seems to be missing. The aim of this work consisted in extracting and analysing the structural features shared by mutations associated with resistance to quinolones in using *S.pneumoniae* as a case study, in order to establish a pipeline for automated analysis of other variations associated with drug resistance.

Mutations were analyzed both manually and computationally in order to highlight interesting features. A machine learning approach was also explored with the aim of developing a tool to predict point mutations associated with drug-resistant phenotypes in unknown sequences.

Firstly, we focused our attention on single point mutations, integrating data from the literature and the results from our computational analysis, then we searched for physico-chemical and structural features common to different mutations occurring in sequence positions associated with antibiotic resistance.

Finally, the features extracted were used to train and test

different machine learning algorithms in order to identify an accurate predictive model.

All the mutations known to be associated with quinolone resistance were manually retrieved both from the literature and from specialized databases such as CARD³⁰, PointFinder²⁶, the NCBI Resistance

Gene Database³⁵ as described in Materials & Methods. Data was annotated and stored in the Quinores3D db. Indeed, in order to compare amino acidic residue types appearing in the susceptible phenotype with those observed in resistant phenotypes, the amino acids at antibiotic-resistant positions in the reference and the mutated sequences from 21 bacterial species resulting in a dataset of 363 residues from different organisms (see Table 4.1). We worked on this set of bacteria because quinolone resistant variations for these organisms are well annotated in the previously mentioned databases.

134 wild type amino acids were labelled as susceptible, 206 substitutions were labelled as resistant, and 19 were labelled as unknown as, despite they were reported in CARD to be associated with resistance, it was not possible from the literature to clearly understand if they were directly causative of resistance or not. They correspond to positions GyrA 61, 70, 91,102,117, ParC 78,99, GyrB 147,398,408,421,432,456 and ParE 447,489 referring to *S. pneumoniae* sequence numbering and were removed from our dataset.

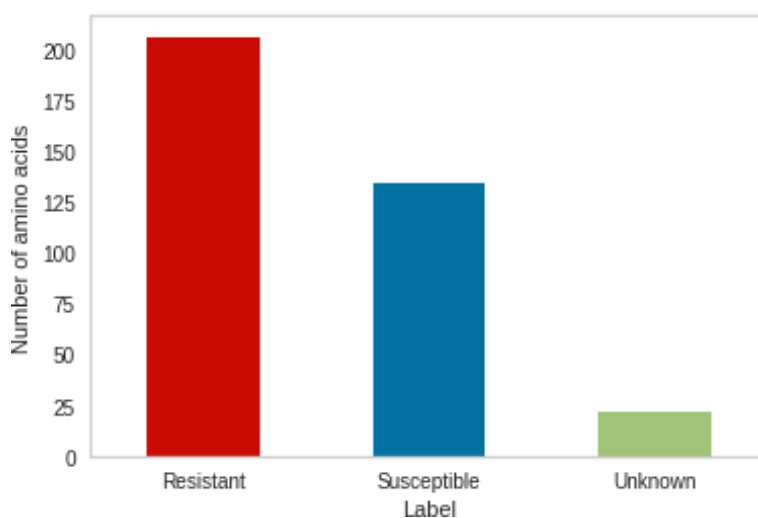


Fig 4.8. Number of amino acids in the dataset according to their category.

Since machine learning methods require numeric variables, each residue was encoded in a set of numeric values, each representing a different physicochemical, biochemical and structural property.

In order to calculate structural properties, a 3D structure was required. So, for each sequence carrying the resistant mutation

we generated a protein model with homology modelling. Similarly, we generated the models from the ‘reference’ sequences, that are the wild type and susceptible sequences. Overall we generated 207 protein models. From each of these proteins we calculate the relative solvent accessibility, electrostatic potential, net charge and charges distribution, distance from drug and magnesium for the amino acid of interest (see Materials & Methods). We decided to generate several models and just not to perform the analysis only on *S. pneumoniae* structures because we are working with proteins from different bacteria, which can have small but significant differences in sequence and structure impacting on the structural properties.

4.5. Cluster analysis of mutations associated with bacterial AR

Mutations associated with drug resistance are located in different regions of proteins, in particular at specific key positions in different subunits of gyrase and topoisomerase IV (see Table 4.1). In order to identify commonalities and differences among key positions we performed a clustering analysis on the residue types occurring at these positions in the reference sequences (i.e. sequences belonging to susceptible bacteria). This made it possible to obtain a picture of the characteristics of the key positions in the wild type gyrase and topoisomerase IV subunits. Then,

for each cluster identified we studied the variations related to AR in the cluster to highlight properties shared by the substitutions and make a comparison with the wild type (and therefore susceptible) amino acids. Finally, we characterize each position in the cluster in terms of the physico- chemical and structural features of residues associated with antibiotic resistance, merging the results of our analysis with what has already been studied from the literature.

For the structural analysis we focused our attention on the following physico- chemical and structural features: relative solvent accessibility, electrostatic potential, net charge and charges distribution, distance from drug and magnesium, residue volume, hydrophobicity and hydrophilicity index, molecular weight of the side chain, polarity index, side chain volume, bulkiness, (see Materials and methods).

Using the k-means clustering method implemented in scikit learn⁴⁶ (see Materials & Methods), five separated clusters were identified.

Cluster	Protein	Position
I	GyrA	E85
	GyrB	E474,E475
	ParC	E83
	ParE	E474,E475
II	GyrA	G79
	ParC	G77
III	GyrA	D80,D85
	GyrB	D435,N473
	ParC	D83

	ParE	D435
IV	GyrA	M81
	ParC	A63
	ParE	P454
V	GyrA	S81/T81
	ParC	S79

Table 4.3. Clusters identified with the k-means algorithm. For each protein in the cluster the positions are reported in the format wild type – position numbering.

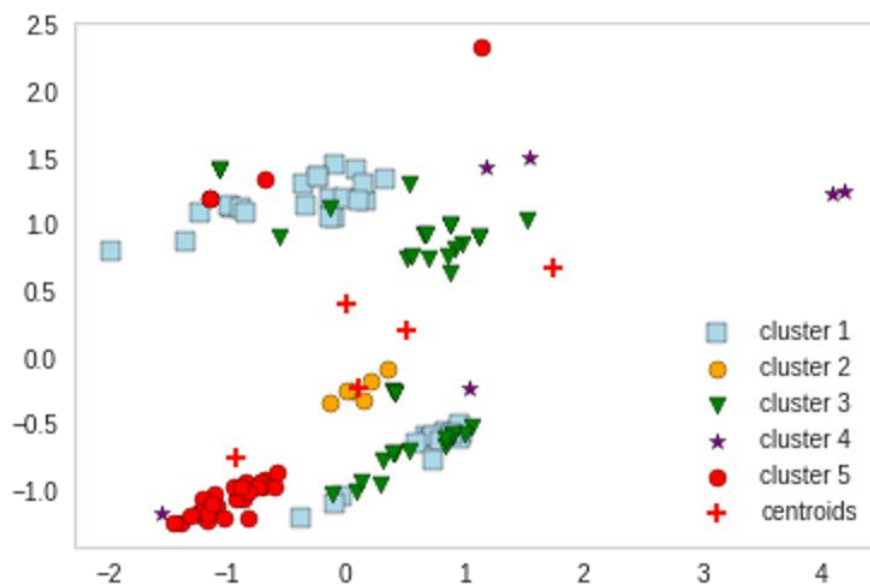


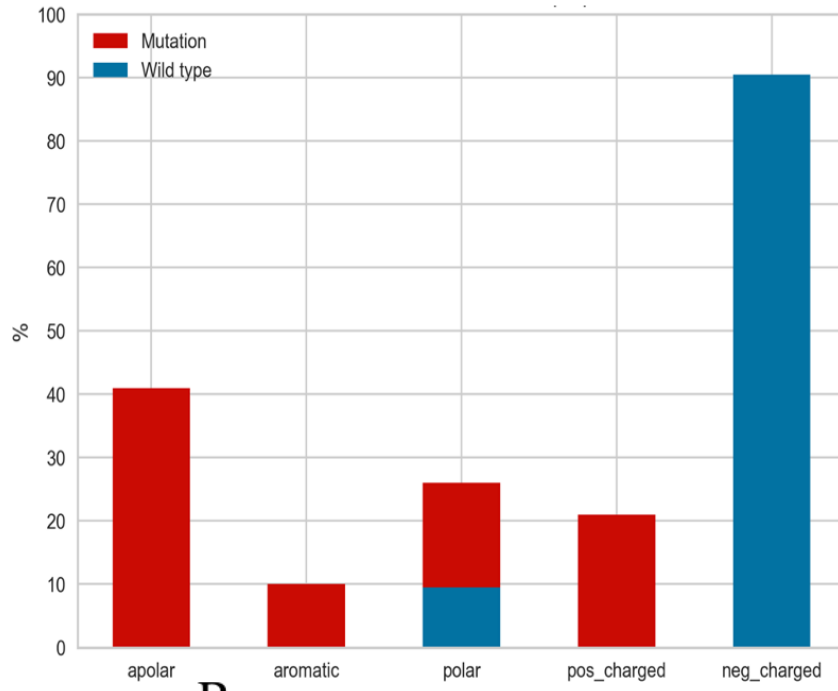
Fig. 4.9. Representation of the five clusters . The centroid (red cross) represents the center of the cluster. We can observe an heterogenic distribution of the cluster number 1 and 3 (blue square and green arrow).

Referring to fig 4.9., we can see how cluster 5 (red dots) and cluster 2 (yellow dots) are well separated, while cluster 1 (blue square) and cluster 3 (green arrow) tend to overlap. Interestingly, the wild types in cluster I and III are all acid and negatively charged. Due to these considerations, we decided to group together the cluster I and cluster III and analyze it as a single cluster.

4.5.1. Cluster I_III characterization

The cluster comprises 164 observations, 100 are marked as resistant and 64 as susceptible. The positions grouped together correspond to positions GyrA 80, GyrA85 / ParC83, GyrB435 / ParE435, GyrB474 / ParE474, GyrB475 / ParE 475. Wild type amino acids are mostly negatively charged (ASP or GLU) except for some organisms in which in position GyrB 474 there is a polar amino acid (GLN).

A



B

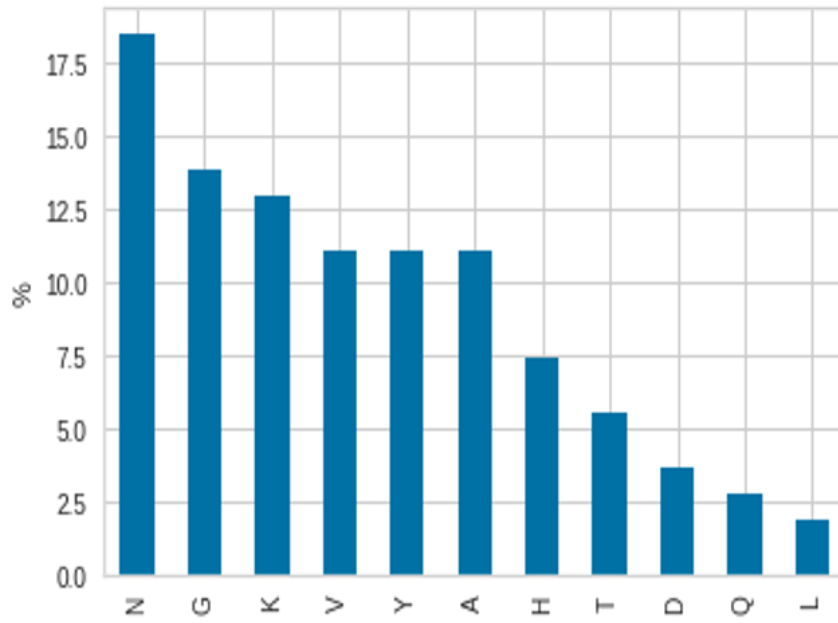


Fig 4.10. A) Percentage of wild type and resistant amino acids according to their chemical properties in cluster I_III; B) Frequency of different amino acid types in the mutated positions in resistant organisms.

As we can see from Figure 4.11, these positions mapped in different regions of the two subunits, with a mean distance from the quinolone of 8 Å, while the mean distance from magnesium can vary from 5 Å to more than 15 Å, due to the fact that the GyrB/ParE subunits are distant from the ion complexed with the drug while GyrA/ParC is very close.

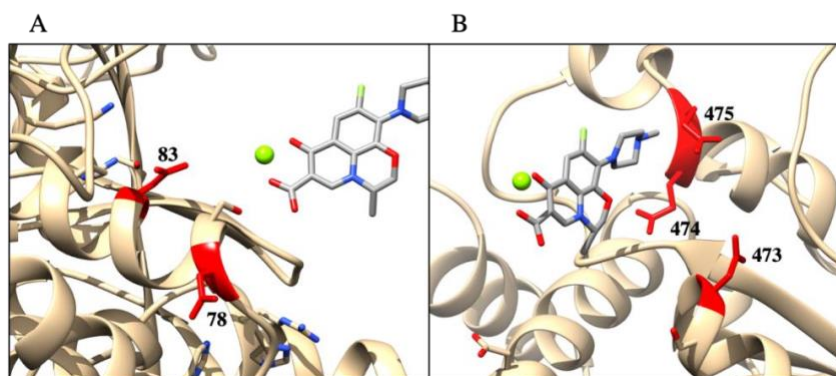


Fig 4.11. Analysis of Cluster I-III: quinolone resistance-associated mutations (A) in the ParC subunit (in red) and (B) the ParE subunit (in red). Magnesium is represented as a sphere (in green) and the quinolone (Levofloxacin) as sticks.

It has been proposed that residues in this cluster interact with the drug by bearing a charge complementary with the positively

charged nitrogens in the quinolone^{22,93} (refer to Figure 1.7), so we explored the effect of charge variations introduced by the mutations. We calculated the electrostatic potential, charge density and net charge value in a radius of 5 Å from the position. Most of the variations lead to the substitution of the negative amino acids with apolar or positive side chains (Figure 4.10), probably perturbing the charge distribution. For the single point mutation analysis we focused our attention on position 83/85, 435, 475. For position 474 and 80 the few data available did not permit a valid statistical analysis, since the p-value calculated for electrostatic potential and charges was above 0.05.

4.5.1.1. Characterization of the ParC 83 / GyrA 85 position

ParC 83 is equivalent to GyrA 85 and it is often associated with quinolone resistance⁷. In *S. pneumoniae* in position 83 the wild type amino acid is an aspartic acid, whereas in the GyrA 85 there is a glutamic acid. The multiple sequence alignment (MSA) in figure 4.12 shows that the position is well conserved among bacteria, while resistant strains are highly variable in that position, as highlighted by the normalized Shannon entropy, which is greater than 0.8 (figure 4.11)

A



<i>parc_bartonella_bacilliformis</i>	VMGK FHPHGD S I Y D A L V R L A
<i>parc_neisseria_gonorrhoeae</i>	I L G K Y H P H G D S S A Y E A M V R M A
<i>parc_acinetobacter_baumannii</i>	V I G K Y H P H G D S A C Y E A L V L M A
<i>gyra_pseudomonas_aeruginosa</i>	V L G K F H P H G D S A C Y E A M V L M A
<i>parc_pseudomonas_aeruginosa</i>	V L G K F H P H G D S A C Y E A M V L M A
<i>parc_haemophilus_parainfluenzae</i>	V L G K F H P H G D S A C Y E A M V L M A
<i>parc_morganella_morgani</i>	V L G K Y H P H G D S A C Y E A M V L M A
<i>parc_klebsiella_pneumoniae</i>	V L G K Y H P H G D S A C Y E A M V L M A
<i>parc_escherichia_coli</i>	V L G K Y H P H G D S A C Y E A M V L M A
<i>parc_salmonella_enterica</i>	V L G K Y H P H G D S A C Y E A M V L M A
<i>gyra_propionobacterium_acnes</i>	VMG K Y H P H G D S A I Y D T L V R L A
<i>gyra_mycobacterium_leprae</i>	T M G N Y H P H G D S I Y D T L V R M A
<i>gyra_mycobacterium_tuberculosis</i>	T M G N Y H P H G D S I Y D S L V R M A
<i>gyra_bartonella_bacilliformis</i>	VMG K F H P H G D S I Y D A L V R M A
<i>gyra_burkholderia_dolosa</i>	V I G K Y H P H G D T A V Y D T I V R M A
<i>gyra_acinetobacter_baumannii</i>	V I G K Y H P H G D S A V Y E T I V R M A
<i>gyra_escherichia_coli</i>	V I G K Y H P H G D S A V Y D T I V R M A
<i>gyra_klebsiella_pneumoniae</i>	V I G K Y H P H G D S A V Y D T I V R M A
<i>gyra_salmonella_enterica</i>	V I G K Y H P H G D S A V Y D T I V R M A
<i>gyra_capnocytophaga_gingivalis</i>	V L G K Y H P H G D S S V Y D T M V R M A
<i>gyra_mycoplasma_genitalium</i>	VM S K F H P H G D M A I Y D T M S R M A
<i>gyra_clostridium_difficile</i>	V I G K Y H P H G D T A V Y D A M V R M A
<i>gyra_staphylococcus_aureus</i>	VMG K Y H P H G D S S I Y E A M V R M A
<i>gyra_streptococcus_pneumoniae</i>	VMG K Y H P H G D S S I Y E A M V R M A
<i>gyra_enterococcus_faecalis</i>	VMG K Y H P H G D S A I Y E S M V R M A
<i>gyra_enterococcus_faecium</i>	VMG K Y H P H G D S A I Y E S M V R M A
<i>parc_mycoplasma_genitalium</i>	I M G K Y H P H G D S S I Y D A I I R M S
<i>parc_mycoplasma_hominis</i>	V I G R Y H P H G D S S I Y E A L V R M A
<i>parc_ureaplasma_urealyticum</i>	V I G K Y H P H G D S S I Y E A M V R M S
<i>parc_staphylococcus_aureus</i>	V I G Q Y H P H G D S S V Y E A M V R L S
<i>parc_streptococcus_pneumoniae</i>	I M G N F H P H G D S S I Y D A M V R M S
<i>parc_enterococcus_faecalis</i>	I M G N Y H P H G D S S I Y E A M V R L S
<i>parc_enterococcus_faecium</i>	I M G N Y H P H G D S S I Y E A M V R M S

Consensus V M G K Y H P H G D S A I Y E A M V R M A

B

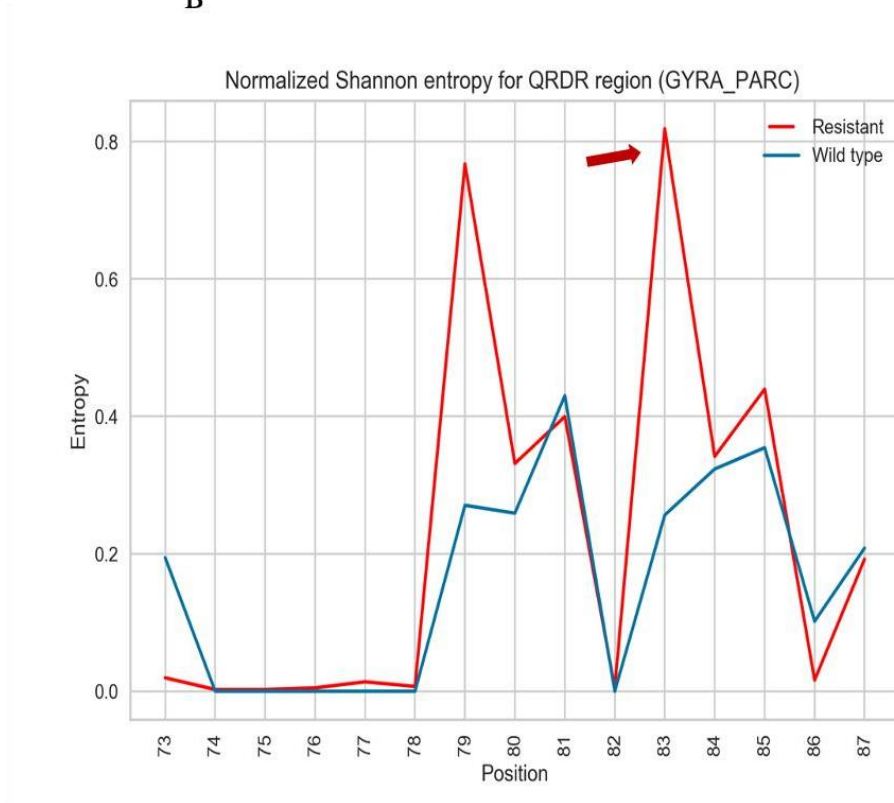
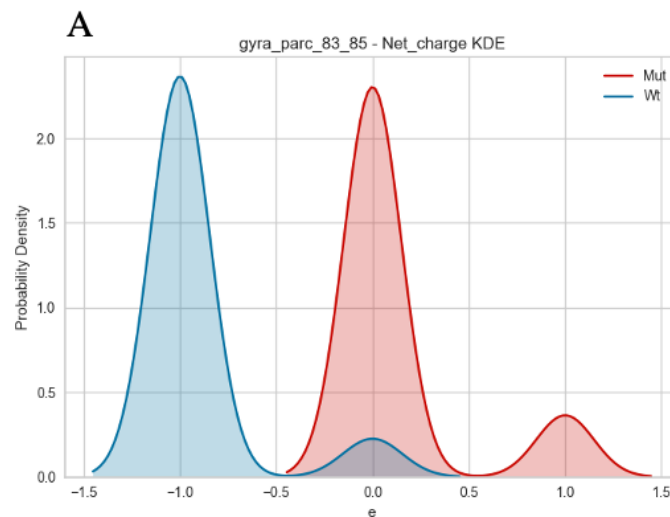


Fig 4.12 A) Multiple sequence alignment for GyrA/ParC reference sequences showing positions from 69 to 79 and colored using the Clustal color scheme. The red arrow indicates ParC 83 / GyrA 85 position. B) Comparison between the normalized Shannon entropy for wild type (blue) and mutated sequences (red) over the QRDR region.

It was proposed that the mechanism of resistance related to this position is due to the disruption of the magnesium water bridge that anchors the drug to the protein^{7,21}. Characterization of the charges calculated on different bacteria reveal that wild type

holds an average net charge of $-1 e$ and a mean density charge of $-11282,60 e/\text{\AA}^3$, complementary with the positive charge of the magnesium. Mutations lead to the loss of the negative charge, with net charge average near to $0 e$ and the density charge becoming positive (calculated as $7498,30 e/\text{\AA}^3$). Figure 4.13 summarizes our results: A) represents the plot of the kernel density estimation (KDE) for the net charge calculated in a radius of 5\AA from position 83/85. A KDE plot is a method for visualizing the distribution of observations in a dataset, analogous to a histogram, using a continuous probability density curve. We can interpret the plot as follows: values from -1.5 to $-0.5 e$ are very likely to be associated with susceptible phenotype (blue area), while resistant residues assume positive values (the spike at 0 and the little bump at 1 , red area). Mutations increase the net charge, which results in a shift from negative to positive density charge (Figure 4.13. B).



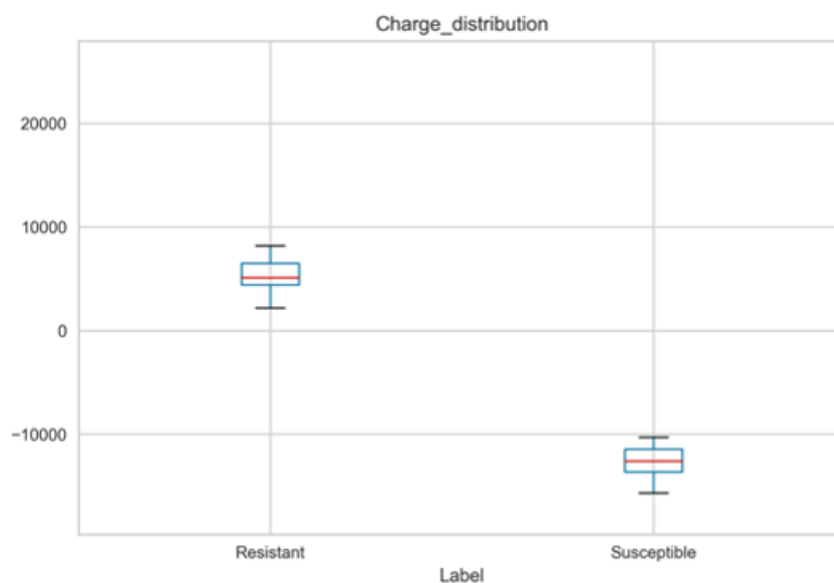


Fig 4.13. A) KDE of the net charge calculated for position 83/85, showing clearly that susceptible residue charge is mostly negative (blue bar) while resistant residues are associated with positive values. B) Box plot of mean values of the charge density for susceptible and resistant residues.

Our findings support the hypothesis that the positive charge introduced by the mutation, and the corresponding loss of negative charge affects quinolone binding by repulsing the divalent metal ion that is chelated by the drug in sensitive organisms, as proposed by Aldred et al²³ (Figure 4.14).

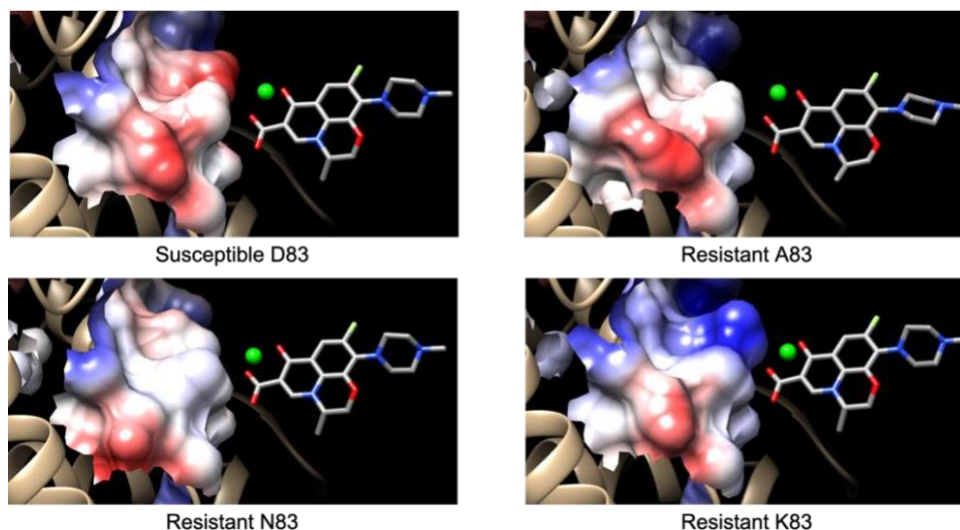


Fig 4.14. Comparison of electrostatic surface for wild type ASP83 (ParC) and 3 resistant mutations (ALA,ASN,LYS). The color of electrostatic potential values ranges from blue (positively charged) to red (negatively charged).

4.5.1.2. Characterization of the ParE 435 / Gyrb 435 position

Unlike the other mutations analyzed so far, this position is located in the ParE and GyrB subunits. The MSA shows a very high conservation among the organisms under study (Fig 4.14). The wild type residue is an aspartic acid that is $\sim 7.5 \text{ \AA}$ distant from the drug and more than 14 \AA from the magnesium; the aspartic acid is supposed to interact with the drug through electrostatic interactions²¹. Mutations replace the negatively charged side chain with positively charged side chains (Lys, Asn)

and apolar residues (Ala), leading to a local increase of positive charges

and a potential disruption of the drug binding.

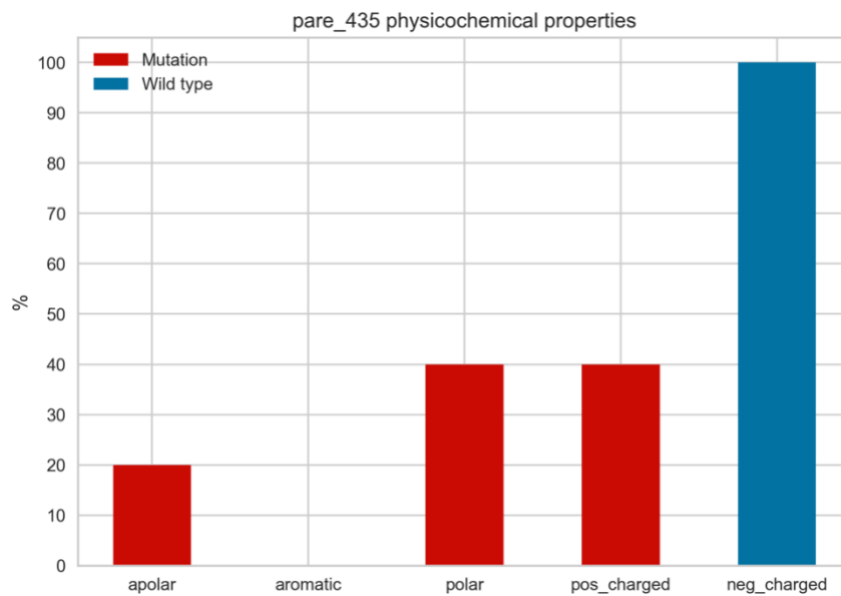


Fig 4.15 Frequency of susceptible amino acids and resistant mutations according to their physicochemical properties.

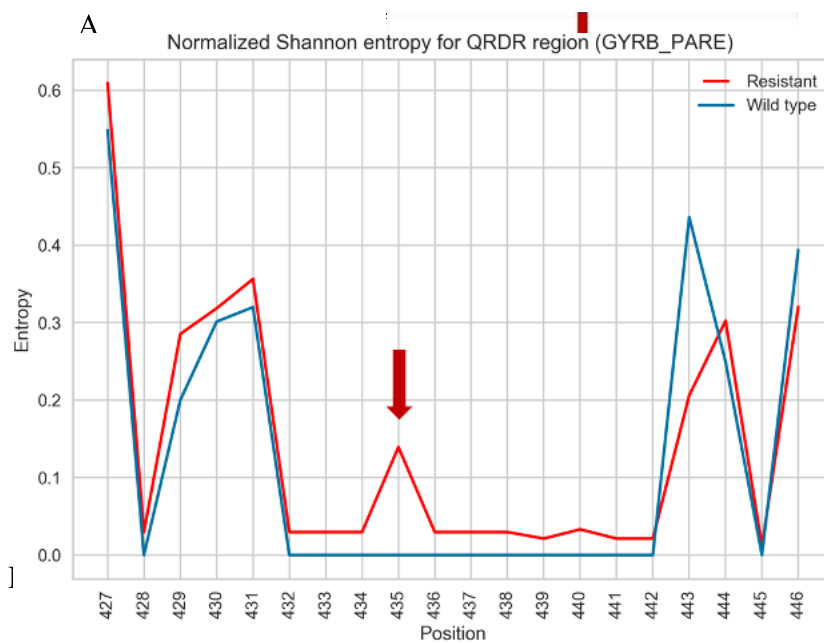


Fig 4.16 A) Multiple sequence alignment for GyrB/ParE reference sequences coloured using the Clustal colour scheme. The red arrow indicates position ParE / GyrB 435. B) Comparison between normalized Shannon entropy for wild type (blue) and mutated sequences (red).

The loss of the negative charge may affect the binding with the positive charge of the fluoroquinolone compound, resulting in the destabilization of the interaction between the drug and the enzyme. The mean net charge for the wild types is $-1.28 e$ and the mean electron density is $-11445 e/\text{\AA}^3$, whereas resistant variations display a mean net charge of $-0.6 e$ and a mean electron density of $-700 e/\text{\AA}^3$ (refer to figure 4.17 C and D).

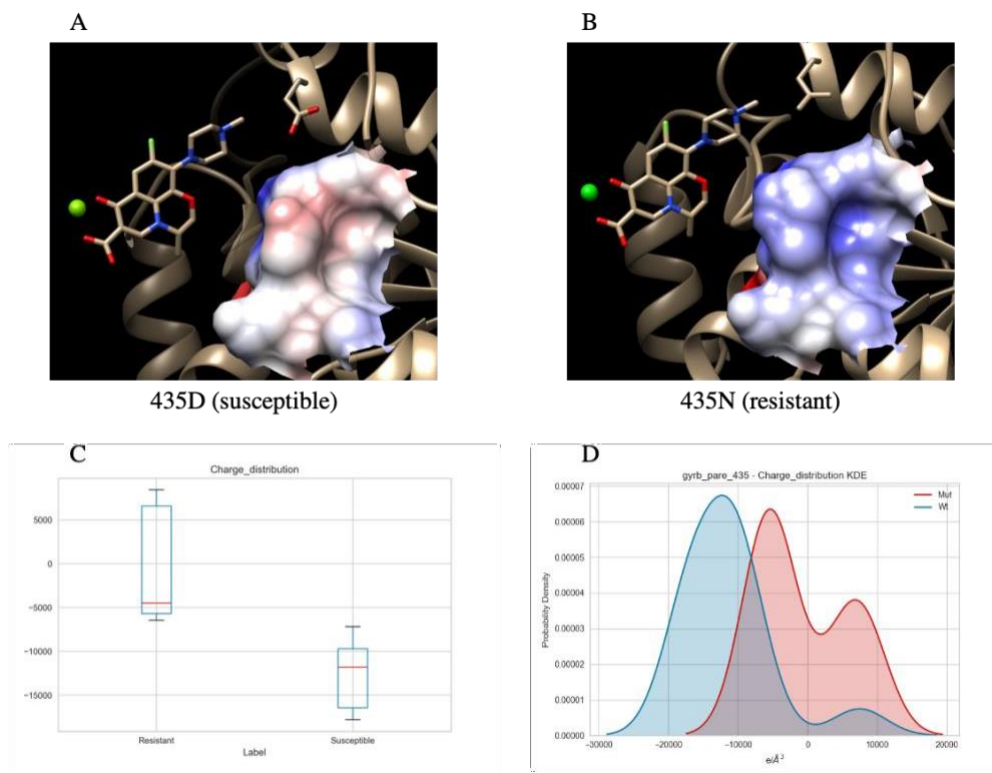


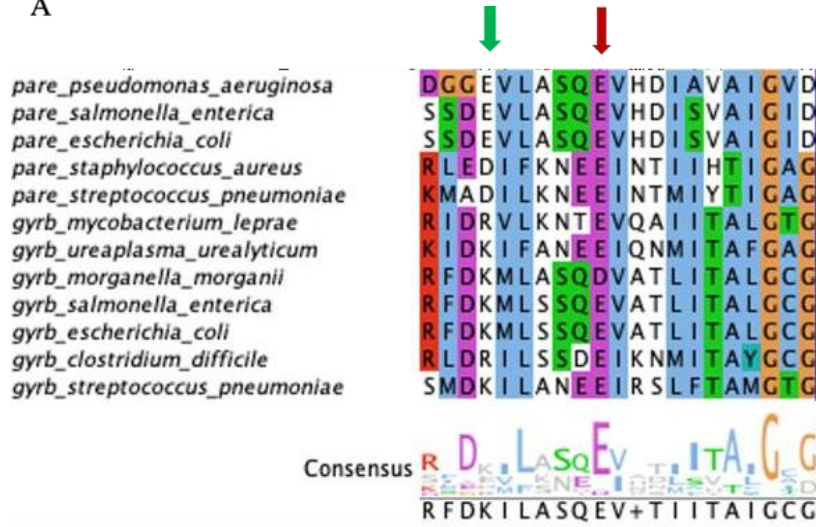
Fig 4.17. A) Electrostatic surface analysis for wild type D435 and B) mutated residue (D435N). The color of the electrostatic potential values ranges from blue (positively charged) to red (negatively charged) C) represent the mean box plot for the charge density and D) the KDE plot for the charge density.

4.5.1.3. Characterization of the ParE/GyrB 475 and ParE/ GyrB 474 positions

Similarly to positions ParC 83 / GyrA 85 and ParE 435 / GyrB 435 , position Par E 475 / GyrB 475 is characterized by the negatively charged glutamic acid, which is extremely conserved

across bacteria (Figure 4.18 A) and is relatively close to quinolone (~5.5 Å).

A



B

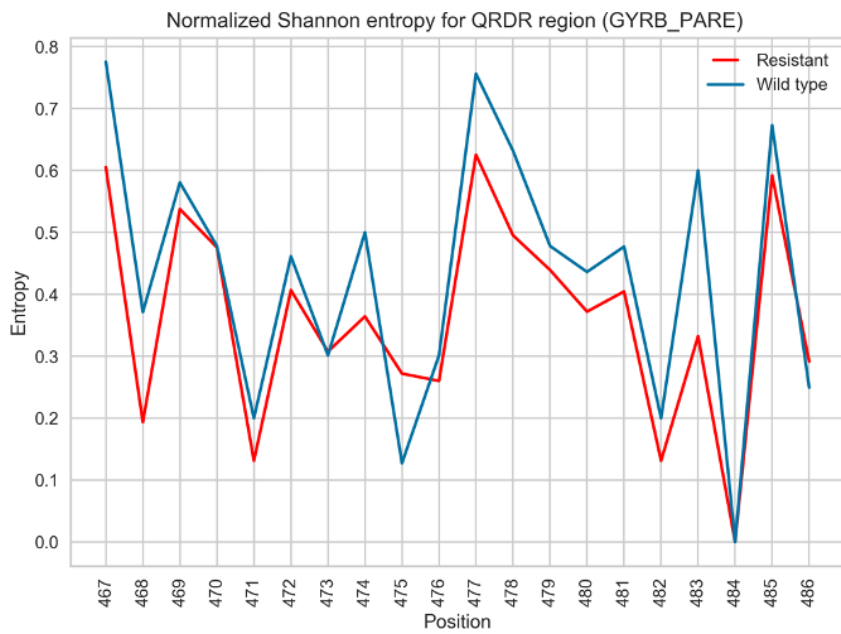
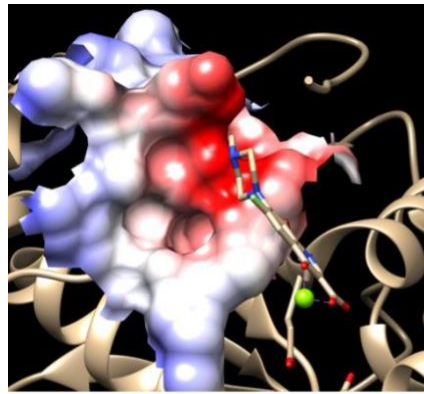


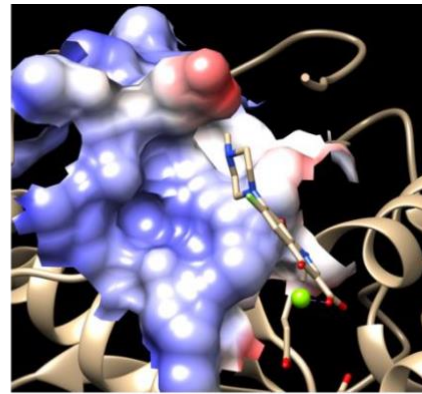
Fig 4.18 A) Multiple sequence alignment of ParE / GyrB reference sequences colored using the Clustal color scheme. The red arrow indicates position ParE / GyrB 475 and green arrow position 474; B) comparison between normalized Shannon entropy in wild type (blue) and mutated sequences (red).

Variations lead to the substitution of the negative side chain with positively charged and apolar side chains (Fig 4.21), resulting in the loss of the interaction with the positively charged drug. In particular, the replacement of the glutamate with a lysine introduces a positive charge and the replacement with an alanine leads to the loss of negative charge.

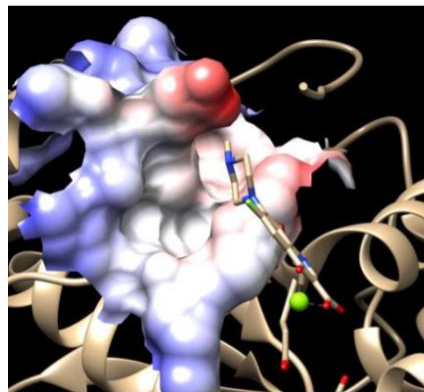
The result is an overall change from a marginally positively charged environment (mean net charge: 0.70 e) to a definitely positively charged one (as mean net charge: 1.70 e), with the introduction of +1 positive charge (Fig 4.19).



Susceptible E475



Resistant K475



Resistant A475

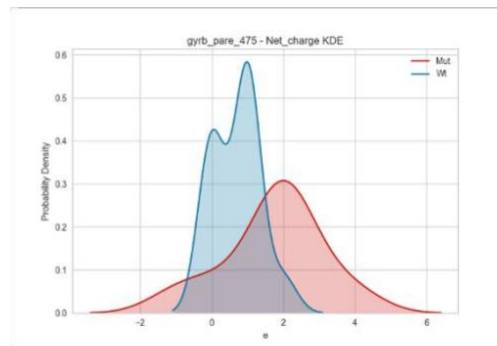


Fig 4.19. A) Comparison of the electrostatic surface of the wild type position (E475) and the mutated position: B) E475K, C) EA475A. The structure is PDB 3RAE chain D. The color of the electrostatic potential values ranges from blue (positively charged) to red (negatively charged). D) Net charge distribution KDE.

Position ParE 474 / GyrB 474 is less conserved than position ParE 475 / GyrB 475 (Figure 4.17). Indeed, the wild type displays several different amino acid types (Fig 4.21), including polar (glutamine and threonine) and negatively charged residues (aspartic and glutamic acid). Mutations introduce aromatic and positively charged side chains, thus modifying the negatively charged environment to a more positive one, possibly destabilizing the interactions between the protein and the antibiotic molecule (Fig 4.20).

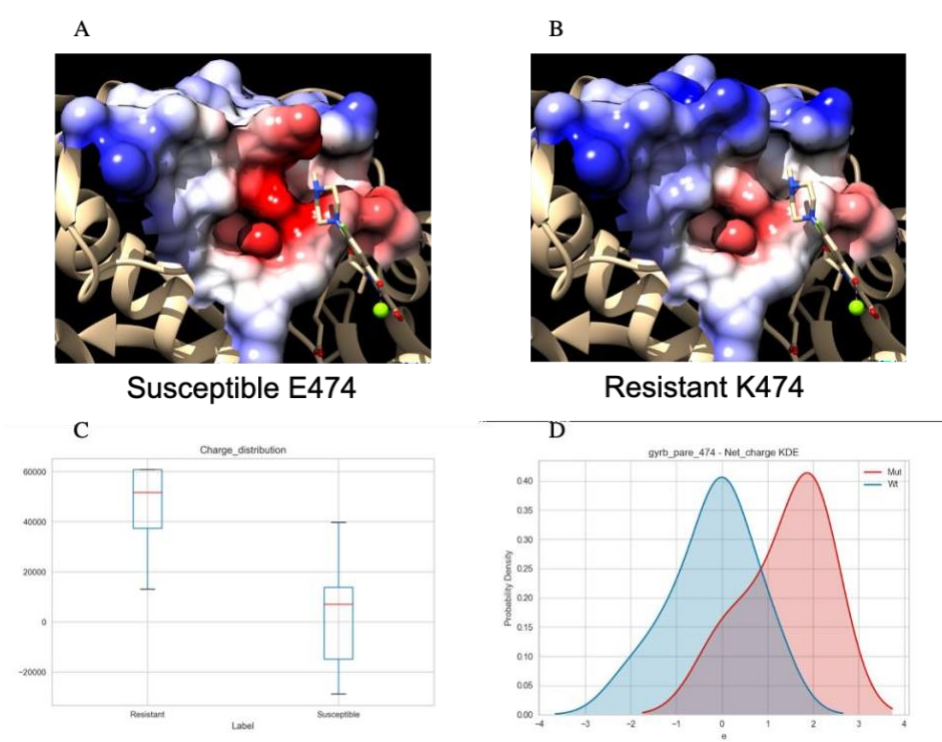


Fig 4.20. Comparison of the electrostatic wild type A) E474 and B) mutated K474. The color of the electrostatic potential values ranges from blue (positively charged) to red (negatively

charged). C and D represent the box plot for the mean charge distribution and the KDE plot

for the net charge distribution. Blue area represents values associated with susceptible residues while red area with resistant.

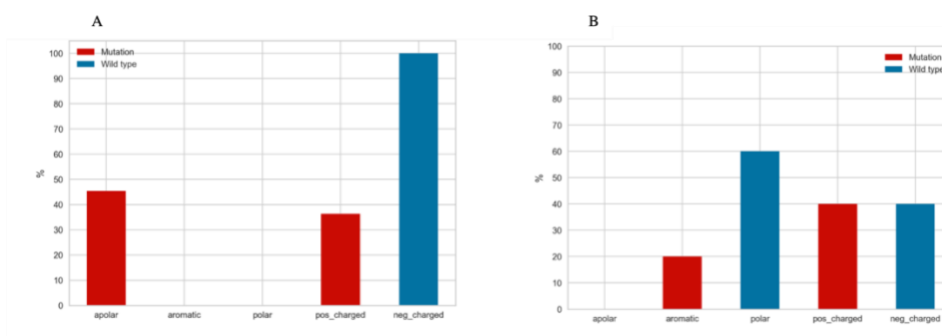


Fig 4.21 Frequency of susceptible amino acids and resistant mutations according to their physicochemical properties for position A) 475 and B) 474.

4.5.1.4. Characterization of the ParC 78 position

ParC 78 is highly conserved and the mutations replace the aspartic acid with apolar and polar amino acids (Fig 4.22). The position is also relatively close to the drug (6.42 Å) and the magnesium (8.5 Å). Based on the net charge change shown in figure 4.23 C, we hypothesize that the loss of a negative charge modifies the overall charge distribution in the region, resulting in the destabilization of the interactions between the mutated residue and the drug.

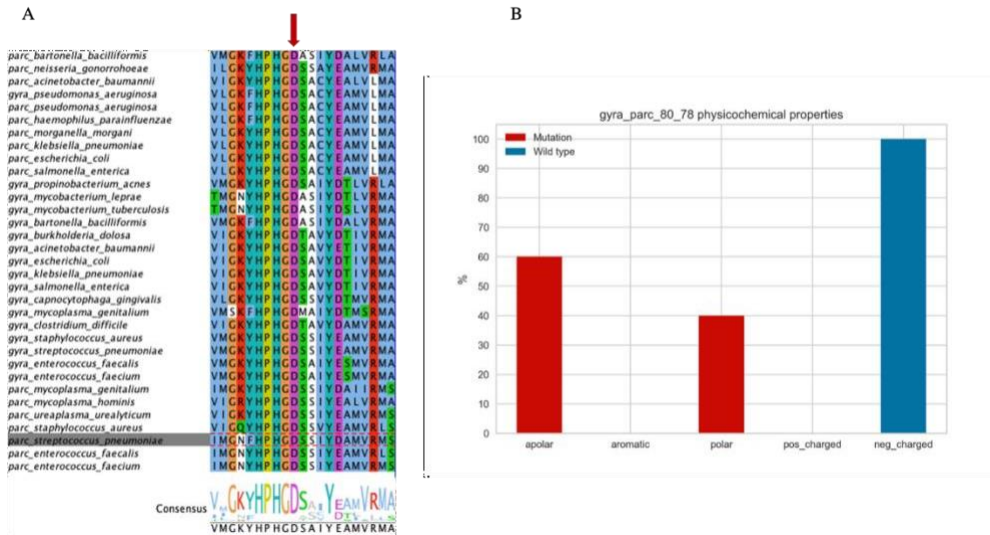
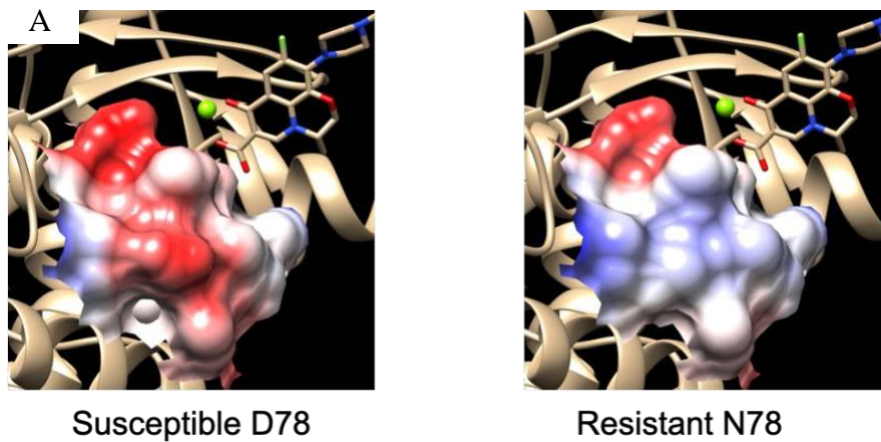


Fig 4.22 A) Multiple sequence alignment of ParC/GyrA reference sequences colored using the Clustal color scheme. The red arrow indicates position ParC 78. B) Frequency of susceptible amino acids and resistant mutations according to their physicochemical properties for position 78.



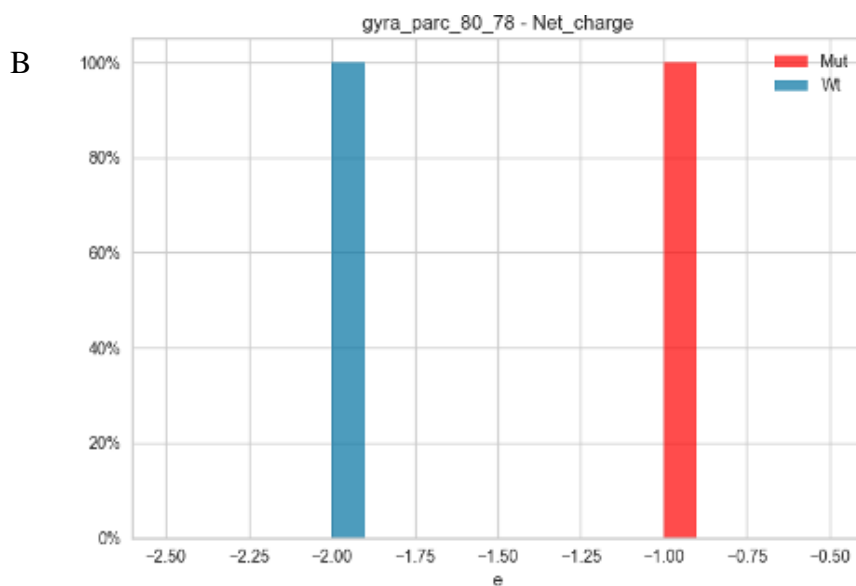


Fig 4.23. A) Comparison of the electrostatic surface for wild type A) ASP78 and B) mutated ASN78. C) The bar plot represents the frequency of net charge values in our dataset, we can clearly observe that susceptible residues are associated with a negative net charge, while resistant residues lead to loss of a negative charge (from $-2.00 e$ to $-1.00 e$).

4.5.2. Cluster II characterization

This cluster corresponds to ParC 77 / GyrA 79 positions, yet mutations associated with AR occurring at position ParC 77 / GyrA 79 are very few and are only found in *Burkholderia dolosa*, *Mycoplasma genitalium*, and *Escherichia coli*. The wild type sequence in position ParC 77 / GyrA 79 presents a very well conserved glycine residue. The only mutations reported in CARD

are changes from glycine to aspartic acid, cysteine and histidine (Fig 4.24).

Similarly to position ParC 78, position ParC 77 / GyrA 79 is relatively close

to the drug (mean distance 5.78 Å) but it does not seem to be directly involved in the binding.

Replacement of a glycine residue with a residue with a bulkier side chain may result in steric hindrance and in modification of the relative solvent accessibility (Box plot Figure 4.25 D), possibly interfering with the serine in position ParC 79 / GyrA 81 which is necessary for drug binding.

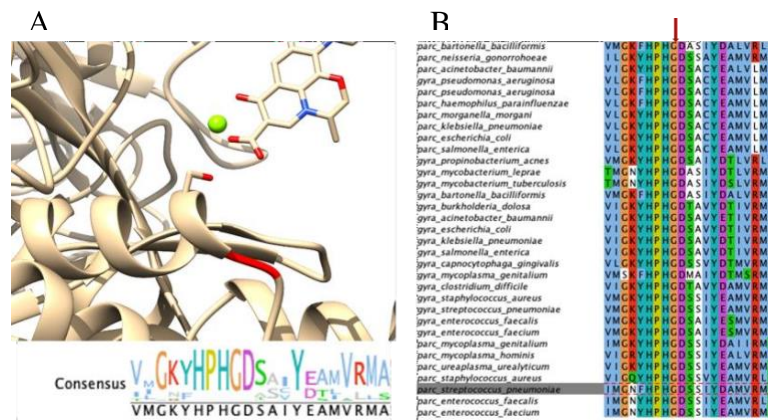


Fig 4.24. A) Map of ParC 77 Gly (in red) on the model, B) the multiple sequence alignment and the logo with the amino acid frequency for the region of interest. The position ParC 77 is indicated by a red arrow.

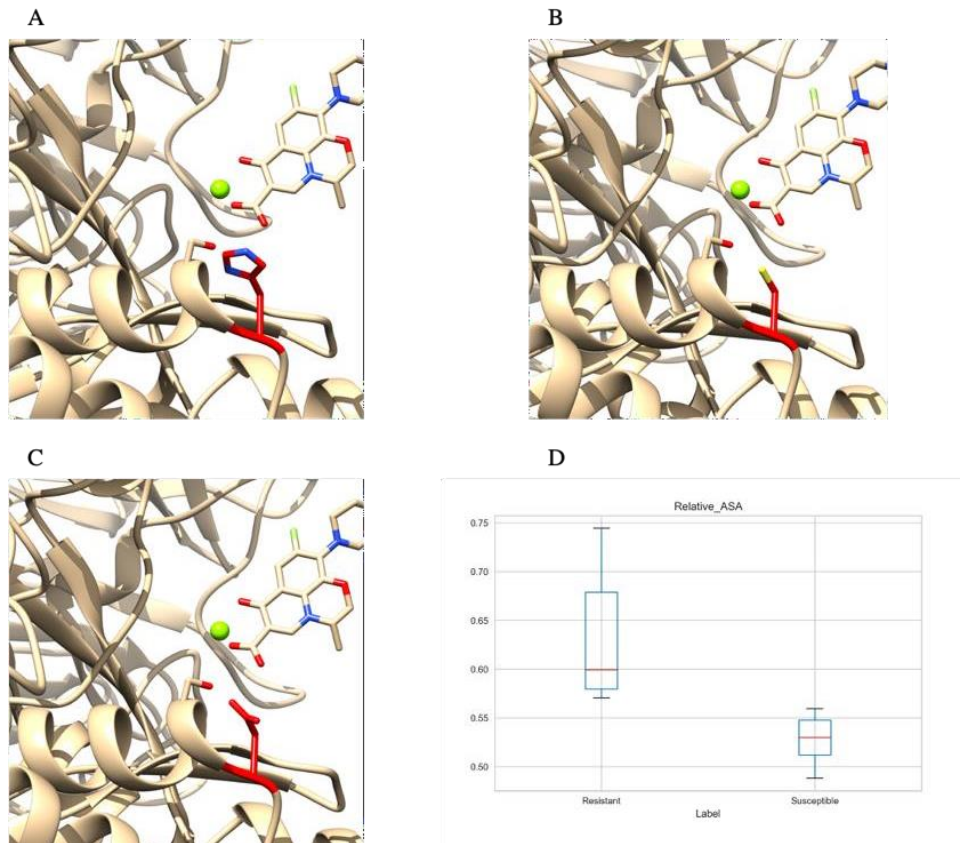


Fig 4.25. Representation of the resistant mutant A) histidine , B) cysteine and C) aspartic acid. In D) it is represented the box plot of the mean values of the relative ASA calculated for resistant and susceptible residues at position 77.

4.5.3. Cluster IV characterization

Identifying a pattern of resistance in this cluster proved difficult due to the heterogeneity of wild type and mutated residues and little data available.

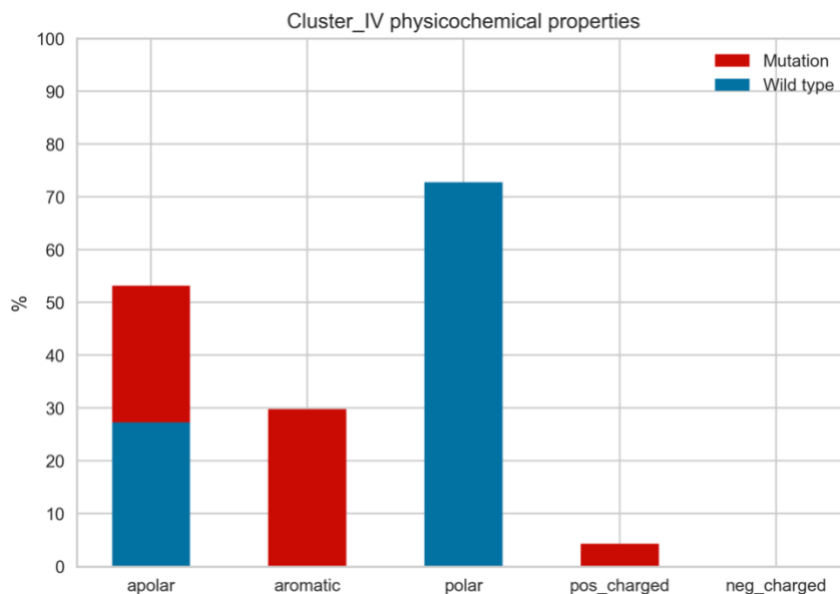


Fig 4.26. Frequency of susceptible amino acids and resistant mutations according to their physicochemical properties.

GyrA M81I was only found in the GyrA subunit of *M. genitalium*, where the reference sequence has a methionine instead of a serine. It was proposed that the wild type enzyme is partially resistant to quinolone and the change of the methionine into an isoleucine residue was associated with moxifloxacin (a quinolone of 4th generation) treatment failure in one patient⁹⁴. Since a polar amino acid is required to bind quinolones, we can hypothesize that apolar amino acids can disrupt the binding, with a mechanism similar to the one described for Cluster II mutations.

ParC A63T and A67S were found in *S. pneumoniae* and *E. coli*, respectively, but they were associated with drug resistance only after the introduction of the above resistant mutations in ParE and GyrB subunits, implying that this mutation alone is not causative of resistance but can contribute to increase it⁸. Moreover, it is distant from the drug, the magnesium and positions ParC 79 and 83 (> than 15 Å). It is unlikely that mutations at this position can interfere with the quinolone binding directly.

ParE P454S was found in *S. pneumoniae*, but even though it has been reported as causative of resistance in CARD, there is no evidence that this mutation can be associated with drug resistance⁹⁵. Also, the distance from the drug (more than 9 Å) or other residues involved in quinolone binding is large enough to suggest that this position cannot interfere with drug binding.

4.5.4. Cluster V characterization

The cluster comprises 84 observations, 60 are marked as resistant and 24 as susceptible. All the observations refer to position 79 of ParC, corresponding to position 81 of GyrA (position numbers refer to *S. pneumoniae* sequence numbering). The position is well conserved among bacteria, with a calculated normalized Shannon entropy below 0.3. Wild type amino acid is generally a serine (74%), followed by alanine (12%), threonine (6%) and

methionine (3%) (Fig 4.27). The mean distance from the drug is 3.2 ($\delta_x 0.12$)Å and 4.4 ($\delta_x 0.12$) Å from magnesium, with δ_x standard error of the mean.

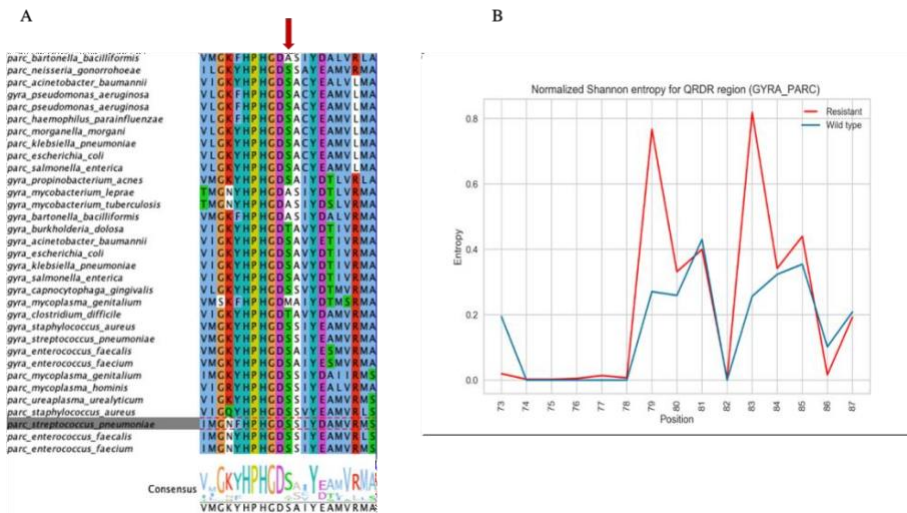
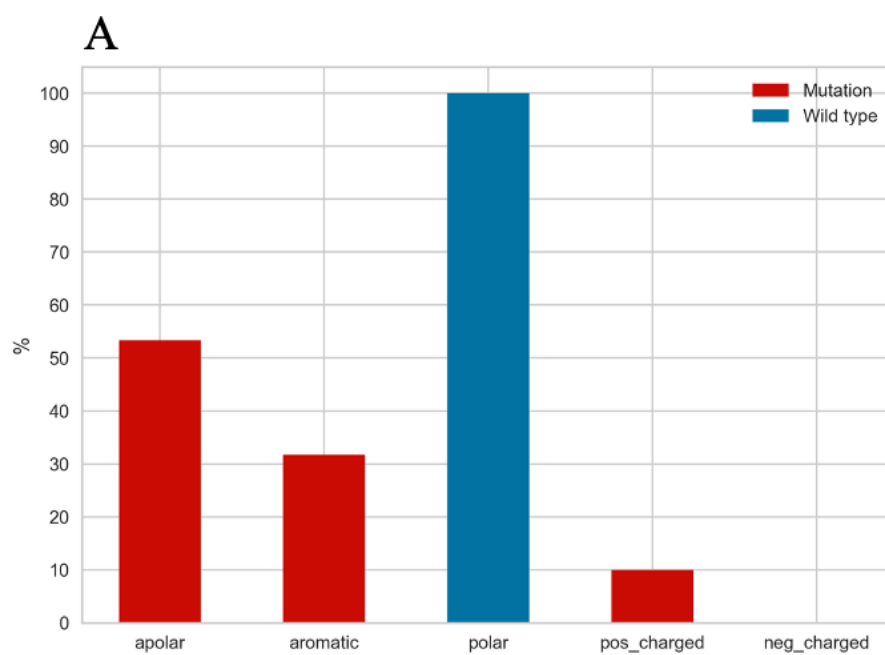


Fig 4.27 a) Multiple sequence alignment for ParC/GyrA reference sequences colored using the Clustal color scheme. The position ParC 79 is indicated by a red arrow; b) comparison between normalized Shannon entropy for the wild type (blue) and mutated sequences (red).

Position ParC 79 /GyrA 81 is a well-known hotspots for quinolone resistance and several substitutions related to quinolone resistance can occur (fig 4.28). Mutations in position ParC 79 / Gyra 81 lead to a change from polar (serine or threonine) to non-polar, aromatic and positive charged residues (Figure 10A) This implies the replacement of the tiny hydrophilic side chain of Ser/Thr with bulky and hydrophobic side chains, as it can be observed in Figure 4.29 from the mean values of

hydrophilicity, hydrophobicity, bulkiness, residue and side chain volume, and polarity of the resistant variations as opposed to wild-type residues.



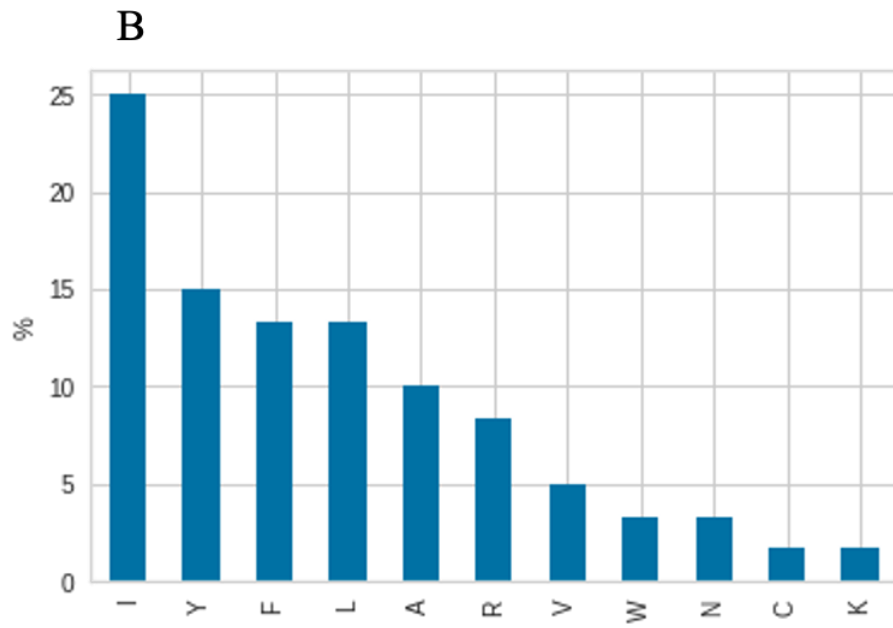


Fig 4.28. A) Percentage of wild type and resistant amino acids according to their chemical properties in cluster V; B) Frequency of different amino acid types in the mutated position ParC 79 / Gyra 81 in resistant organisms.

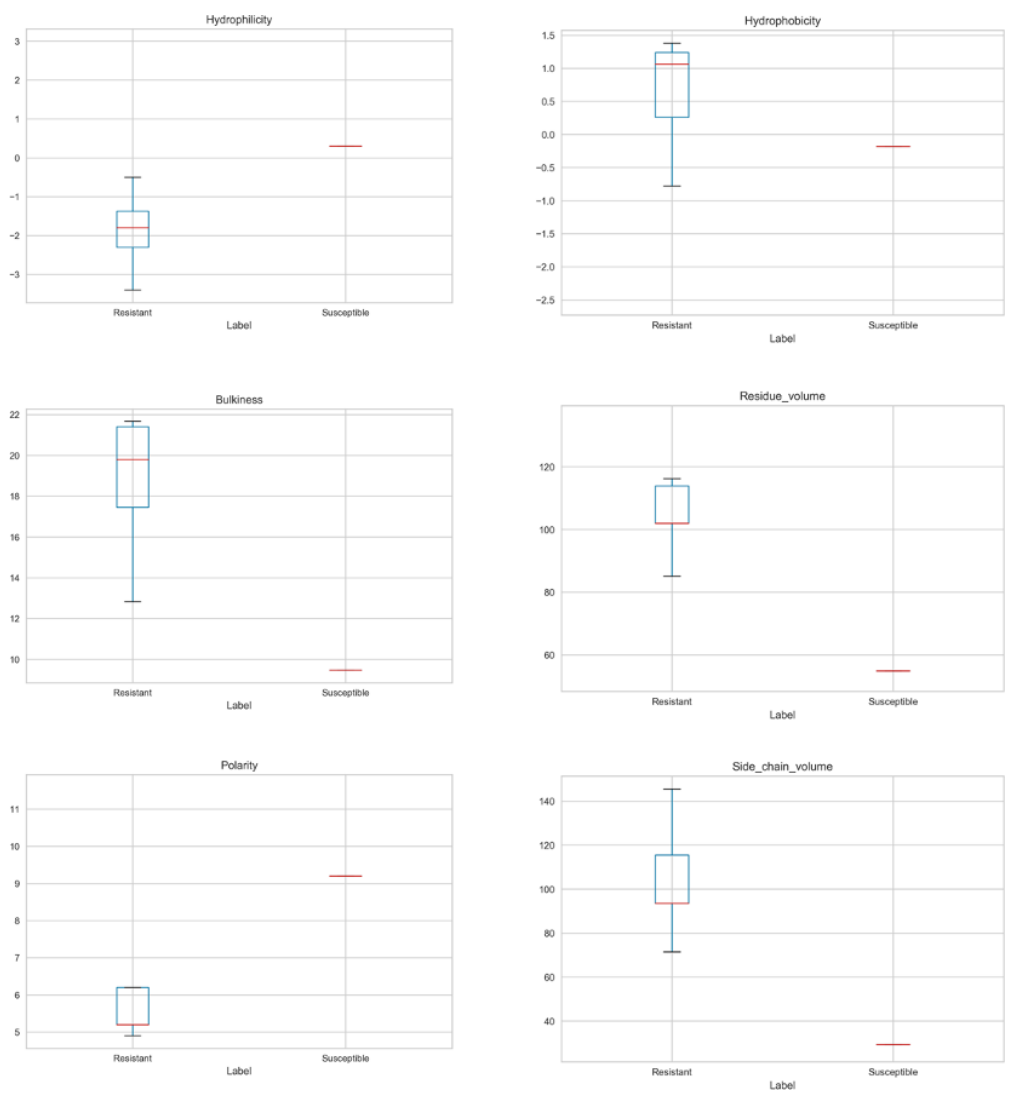


Fig 4.29. Box plot of the mean values representing the AA index features for the resistant residues vs the susceptible residues.

4.5.4.1. Characterization of the ParC 79/ GyrA 81 position

Referring to Figure 1.7, the serine residue forms two hydrogen bonds with the carboxyl group of the fluoroquinolone.

Structural analysis reveals that mutated residues introduce hydrophobic bulky chains resulting in steric hindrance and in the loss of the hydrogen bonds between the -OH group of the serine and the quinolone (see Figure 4.31). The proximity with the magnesium suggests also that mutations can perturb the magnesium - water bridge necessary for drug binding.

Inspired by the work of Kyte and Doolittle⁹⁶, we decided to compare the grand average of hydropathy (GRAVY)⁹⁶ for both the susceptible and resistant sequences. We generate a short region of residues in a range of 5 Å from GyrA81/ParC79 position and calculated the hydropath index of each residue divided by the length of the sequence. Results show an increase in hydrophobicity, with a shift of the mean value from 0.05 (δ_x 0.01) for the susceptible wild type residues to 0.2 (δ_x 0.04) for the resistant ones (refer to Figure 4.30).

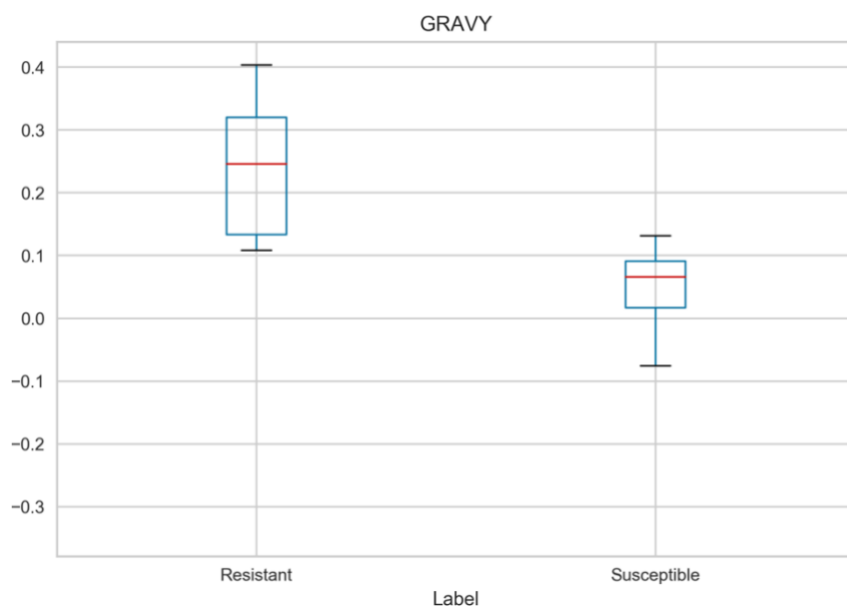


Figure 4.30. Box plot of the mean values of the average of hydropathy calculated in a range of 5 Å from the position.

Since some fluoroquinolones are hydrophilic⁹⁷ (like ciprofloxacin) we can suppose that the shift from a hydrophilic environment to a hydrophobic one can interfere with quinolone binding. Moreover, the bulkier hydrophobic groups result in a physical obstruction for quinolones to enter into 'the quinolone pocket' as proposed by Yoshida et al⁹⁸.

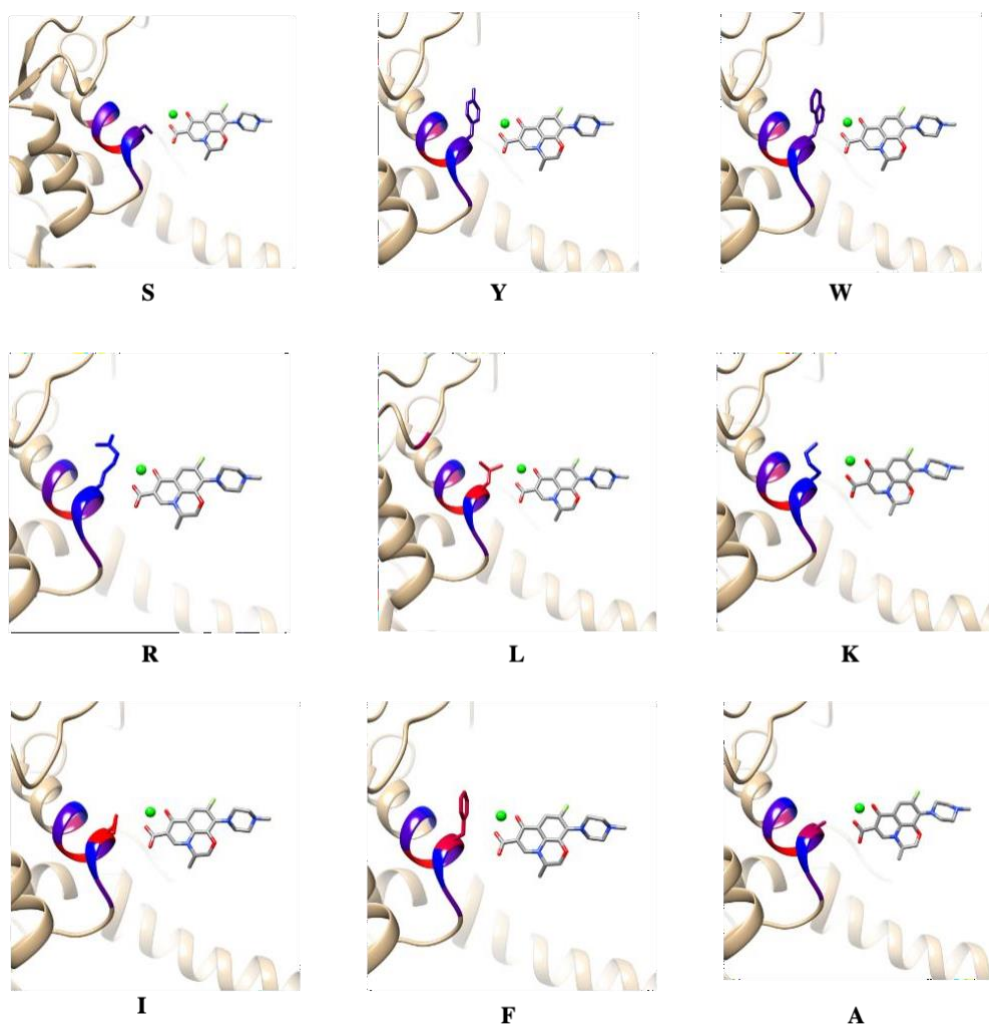


Fig 4.31. Representation of wild type susceptible residue (S) and all possible mutations associated with drug resistance colored according to their hydrophobicity. Color range is from blue (high hydrophilicity) to red (high hydrophobicity). Magnesium is represented as a green sphere and the quinolone (Levofloxacin) as sticks.

4.6. Semi-automatized pipeline for the identification of point mutations in *S. pneumoniae*

ParC_finder was the first method developed for the identification of point mutations associated with quinolone resistance. It is a semi-automatized pipeline which can be used to identify and map variations onto the ParC subunit of *S. pneumoniae* starting from an assembled bacterial genome.

The pipeline has been developed in Python.

Figure 4. describe how *ParC_finder* works. We applied this method on a set of 46 complete genomes downloaded from NCBI Microbial Genomes. The nucleotide sequence is extracted from the genome with Blastn and converted in amino acidic sequence with Blastp. All the protein sequences are collected and aligned with MUSCLE to generate a MSA that is parsed by a script written in Biopython. This script loops through the rows of the alignment and extracts all the mutations, comparing each protein sequence with the ParC R6 susceptible sequence. All the variations are annotated and compared with a list of mutations known to be associated with quinolone resistance in *S. pneumoniae* and mapped on the ParC model (See Result section 4.1.1). The outputs are: a tabular file reporting all the variations, the position, if it is known or not as determinant of antibiotic resistance; Chimera sessions generated by the software, which

contains the model and the residues found mapped on the structure.

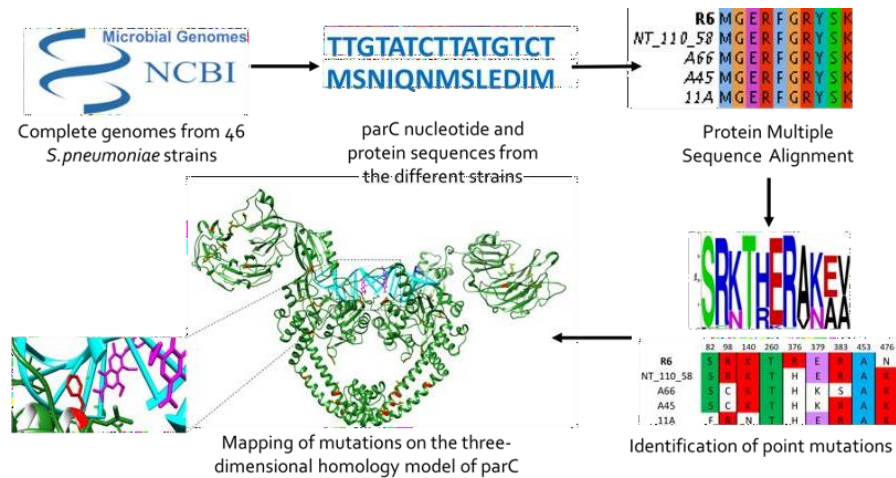


Fig 4.32 Steps of the semi-automated procedure applied on the genomes from several *S. pneumoniae* strains.

4.7. *Quinores3D_pred*: machine learning algorithms for the prediction of point mutations

Several programs for the study of the antibiotic resistance phenomenon have been released in past years. Many of them are able to identify point mutations related to antibiotic resistance like RGI³⁰, PointFinder²⁶ or NCBI- AMRFinder³⁵. All these programs use BLAST to compare the sequence of interest against a reference sequence in order to highlight variations.

However, the identification of antimicrobial resistance-conferring chromosomal mutations often is available for only a limited set of

pathogenic microorganisms, and in case of identification of new variations not identified before, no inference about their role in drug resistance is given. As far as we know, machine learning has been used to detect AMR genes, but not applied on point mutations associated with antimicrobial resistance. Our aim was to

develop a new algorithm for the identification of quinolone resistant variations occurring in the topoisomerases proteins.

Our problem can be considered as a binary classification: given a point mutation (or better, given an amino acid in our protein) the algorithm must predict if it is associated with drug resistant (label 'Resistant' or 1) or not (label 'Susceptible' or 0). We developed two models, one based on Random Forest⁶⁰ and one on Neural Network⁵⁵.

Here, we present preliminary results, which confirm that the approaches are promising and deserve further exploration and development.

4.7.1. Training, test and validation set

Data from clustering analysis (see Results, section 4.5) consists of a set of residues occurring in topoisomerase sequence positions associated with drug resistance and a set of residues present in the same positions in "reference" sequences, namely sequences belonging to bacterial strains that are susceptible to the drug. It is worth noticing that the proportion between the two classes is slightly unbalanced, with the resistant class representing 63% of

the observation. In order to avoid errors during the training phase, we oversampled the susceptible class, reaching a proportion of 50% and 50% for the two categories.

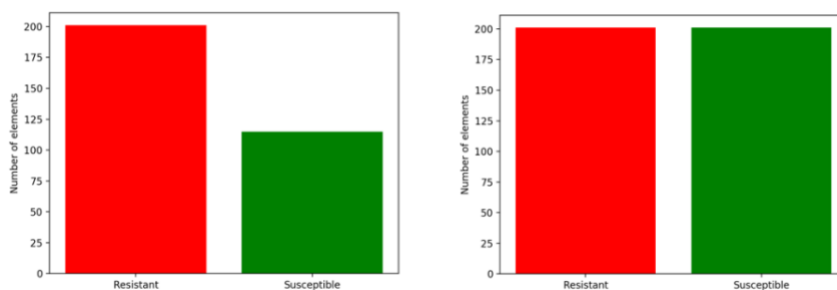


Fig 4.33. Proportion of the two classes before (left) and after (right) oversampling of the minor represented class.

The dataset was normalized in order to have the input variables in the range of 0-1, then it was splitted in a training (70 %) and a validation (30 %) set . The test set was made up of 2123 unseen drug resistant variants and 592 susceptible (wild type) residues, selected from the CARD database and from UniprotKB/Swiss-prot. See table 4.4 for an overview of the data used.

	Training set		Validation set		Test set	
	wild-type	mutated	wild-type	mutated	wild-type	mutated
CARD	0	56	0	9	0	2103
Uniprot	85	111	30	25	592	20

Table 4.4. Overview of the data used for training, validation and test set according to their source (CARD or Uniprot). Wild-type represent the susceptible residues, while mutated the resistant ones.

4.7.2. Feature extraction

For the machine learning problem, both structural and sequence features were extracted. Structural features were: electrostatic interactions, relative ASA, and distance from drug and from magnesium. For the sequence features, we extracted from the AAindex⁹⁰ information about the hydrophathy index, the volume and size of side chain, stability, polarity and free solvation energy for each amino acid.

In order to compute distances and charge values, for each observation a three- dimensional model was generated with homology modelling (see Results section 4.1 and Materials and Methods, section 3.3).

4.7.3. Machine learning models

Seven algorithms commonly employed in machine learning were compared using the accuracy on the training set as a method to choose the one that best fitted with our data. Each model was evaluated using a 5-fold cross validation, and the mean accuracy for each model was collected. Among the algorithms, Random

Forest (RF) reached the highest mean accuracy (94%) on training data (Table 4.5).

Algorithm	Mean accuracy	STD
LR	0,90	0,02
LDA	0,91	0,03
KNN	0,92	0,02
CART	0,95	0,03
NB	0,87	0,03
SVM	0,82	0,03
RF	0,95	0,02

Table 4.5 Algorithm comparison with mean accuracy and standard deviation(STD). LR: Logistic regression, LDA: Linear discriminant analysis, KNN: k-nearest neighbors algorithm, CART: Gaussian Naive Bayes, SVM: Support vector machine, RF: Random forest⁵⁹.

We also explored the possibility to use deep learning for our classification, choosing a densely connected network as neural network model. Our Neural network (ANN) is composed of 2 hidden dense layers with 16 and 8 units respectively, and an output layer composed of a single unit. A dropout layer was added to avoid overfitting.

The performance for both the Random Forest and the neural network was evaluated with a 5-fold cross-validation. The models with the highest accuracy (one for RF and one for ANN) were chosen and tested against the validation set. To

quantify model performance metrics such as accuracy, confusion matrix, precision, recall, F1-score, and Area under ROC Curve were adopted.

RF achieved an accuracy of 93 % on the validation set with an average accuracy of 94% on the 5 fold cross-validation.

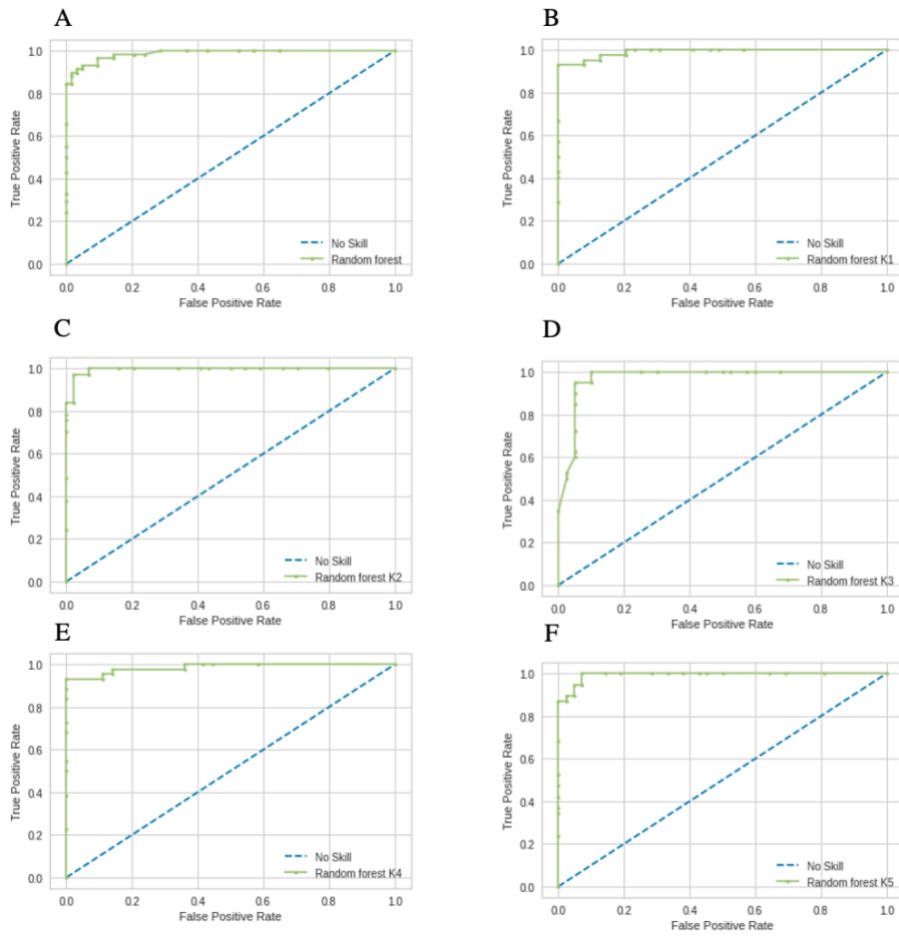
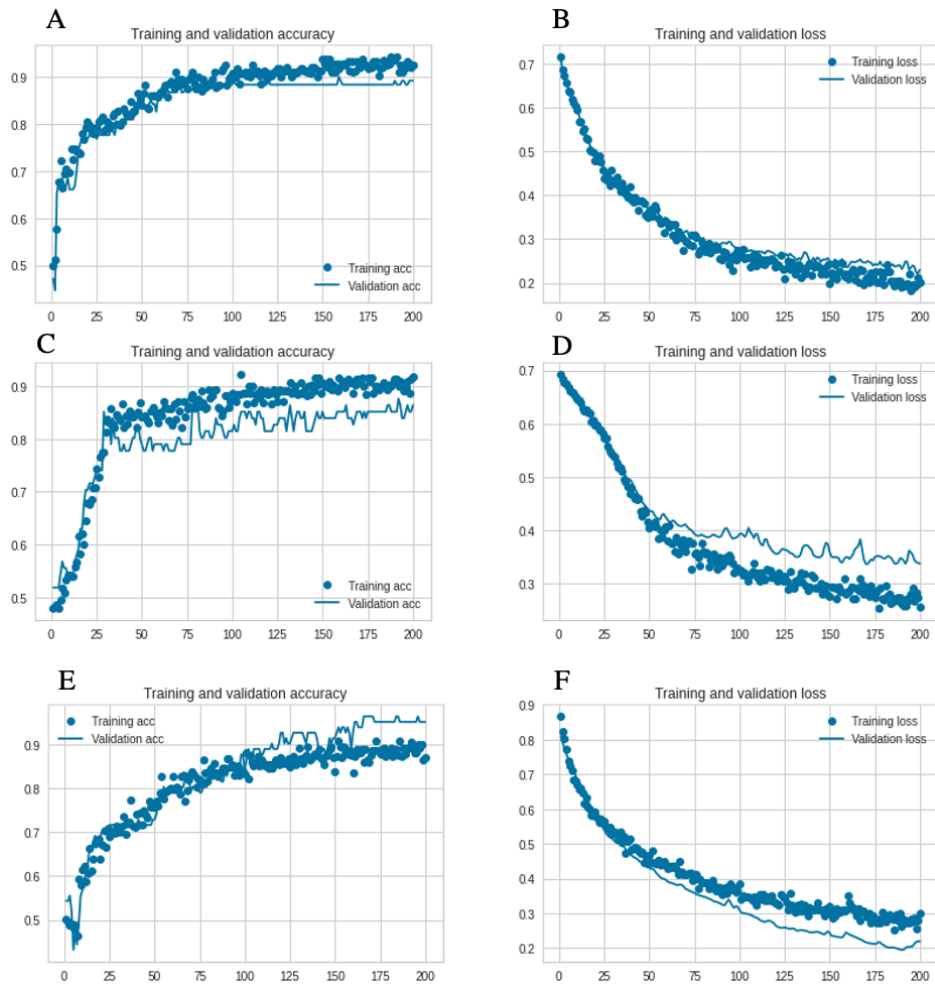


Fig 4.34. ROC curve for the Random Forest models trained on the (A) training/validation set and (B-F) on the 5 fold cross-validation sets.

ANN performed with an accuracy of 89% on the validation set and on the cross validation the mean accuracy was 90%.



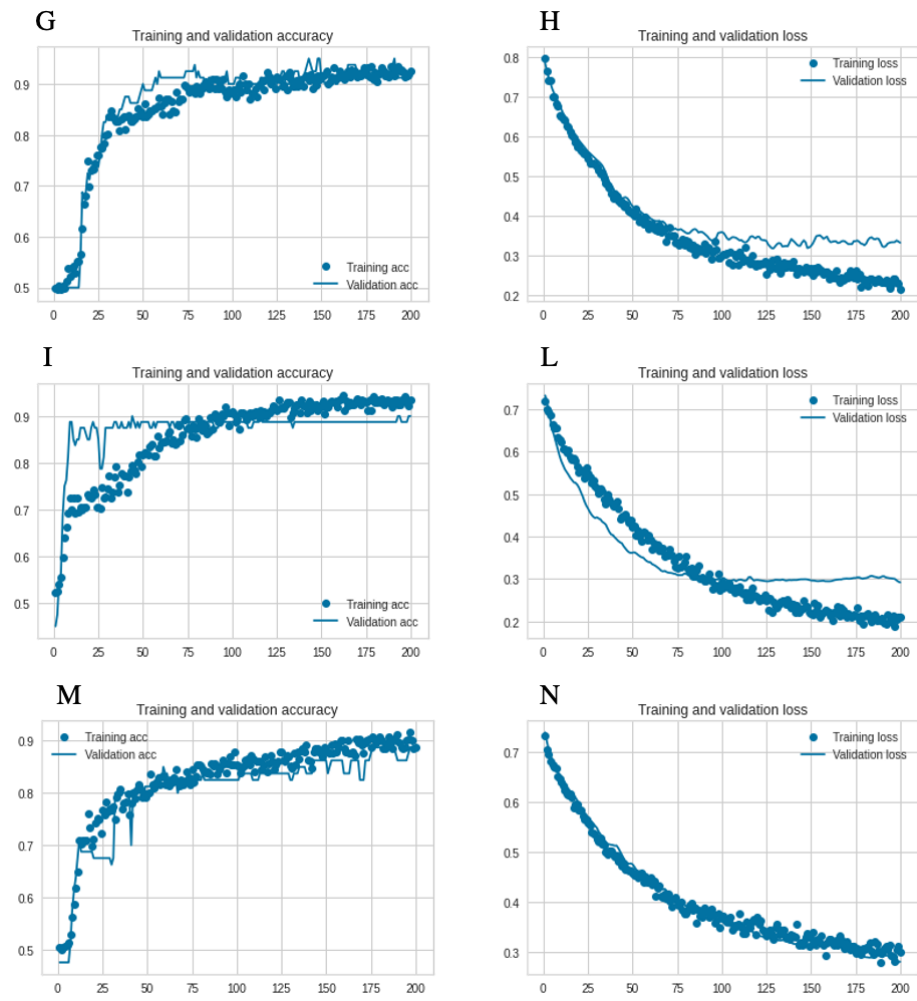


Fig 4.35. These plots represent the metrics for the ANN models trained on the training/validation set (A - B) and on the 5 fold cross-validation (C-N).

Dots represent the accuracy on the training set while lines represent the accuracy on the validation set (A,C,E,G,I,M). Similar for the loss, dots indicate the loss on the training and lines on the validation. The model performs well, yet we can

observe a little overfitting (D and L) which had a negative impact on the accuracy of the classifier (C and M).

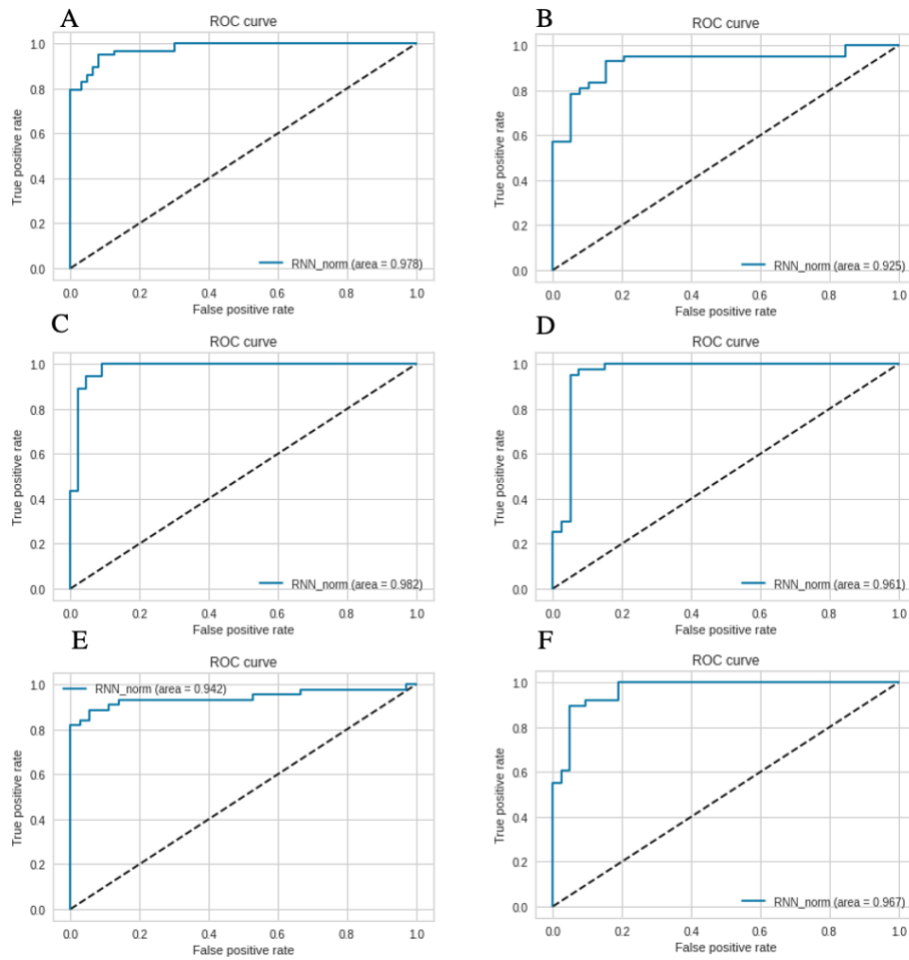


Fig 4.36. ROC curve for the random forest models trained on the training/validation set (A) and on the (B-F) 5 fold cross-validation.

Comparing the two algorithms, the RF performed better than the neural network, with a mean of 94 % of accurate predictions against the 90% of the ANN with the training/validation set.

Then, the two models were evaluated with the test set. Since this set is unbalanced (2123 resistant point mutations associated with resistance and 592 non-resistant), 500 amino acids for both the two classes were picked randomly, in order to generate a random balanced set of 1000 amino acids. RF achieved an accuracy of 94 % on the test, with 49.9% labelled correctly as resistant (true positives), 44.9% as non-resistant (true negatives), 0.1% (1 residue) as false negative and 5.1 % as false positive. ANN performed worst, with an accuracy of 88% and a considerable number of residues predicted as false positive (11.4 %). Refer to the confusion matrices in Figure 4.37.

	RF	ANN
Accuracy	0,95	0,88
Precision	0,91	0,81
Recall	1,00	1,00
F1 score	0,95	0,90

Table 4.6 Classification metrics for test set, comparison between random forest model(RF) and neural network

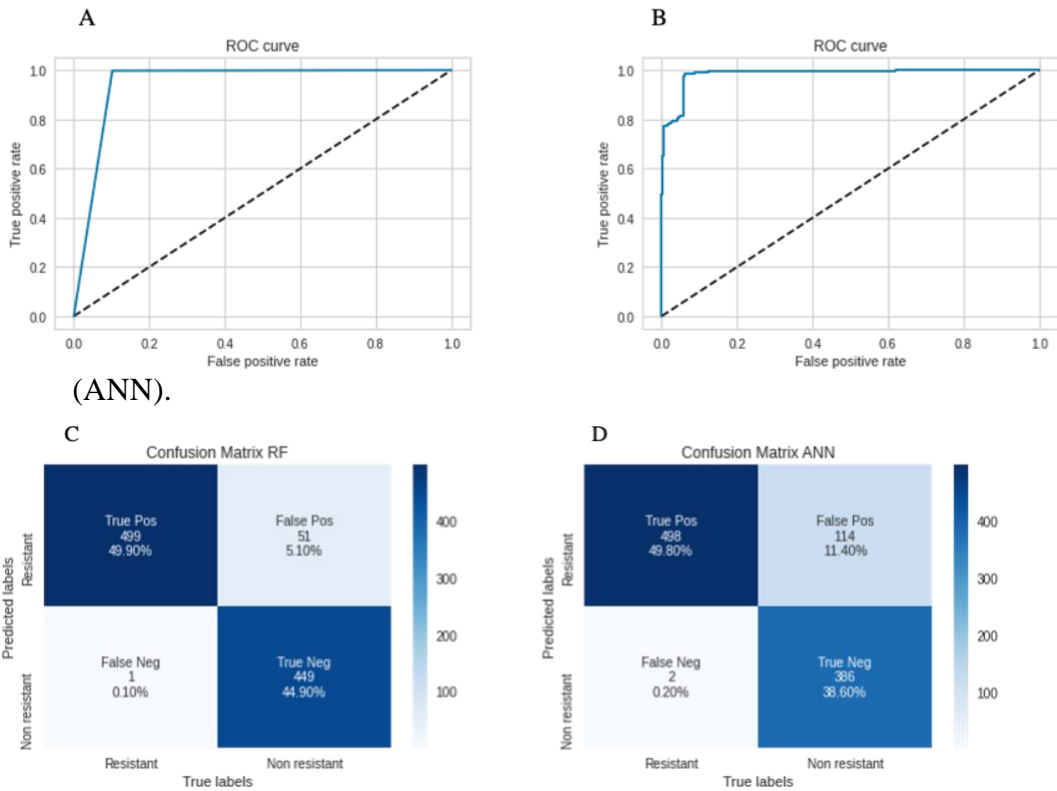


Fig 4.37. ROC curve for the A) Random Forest, B) Neural Network models trained on the test set and confusion matrices for C) RF and D) ANN classifiers.

5. Conclusions

With an estimated 300 million premature deaths and a loss of \$100 trillion, antibiotic resistance will become the first cause of mortality in the world by 2050, surpassing diseases like cancer and diabetes⁹⁹.

Several strategies have been developed to deal with this challenge, including improving the prescription of good practices, optimizing therapeutic regimens, preventing transmission of bacterial infections and improving diagnosis and diagnostic tools¹⁰⁰. But these actions must be taken as soon as possible, in order to avoid an *“unthinkable scenario where antibiotics no longer work and we are cast back into the dark ages of medicine”*⁹⁹.

Given the global reach of the phenomenon, this PhD project focused on the characterization of the molecular mechanisms of antibiotic resistance (AR) caused by the appearance of point mutations and the development of computational approaches for the study and inference of “resistant mutations”, i.e. sequence variations associated with AR to quinolones.

1. Structural characterization of point mutations related to quinolone resistance in S. pneumoniae and other bacteria.

Resistance can arise due to different mechanisms and several classes of antibiotics are affected. We concentrated our efforts

specifically on quinolones as resistance to this new class of compounds is emerging rapidly and the picture of the molecular mechanisms underlying resistance is not yet complete.

Point mutations represent one of the major mechanisms of AR to quinolones. Although much information can be found in the literature, most scientific research deals with the identification of such mutations in bacteria rather than

analyzing and describing in depth the effects on drug binding of amino acid substitutions.

S. pneumoniae was chosen as a case study due to its importance in clinical practice: indeed, it is the fourth cause of pneumonia in the world, and its ability to raise meningitis and fatal sepsis is a threat especially for the elderly and immunosuppressed people.

We presented an exhaustive structural analysis of these variations in the topoisomerase II enzymes using statistical approaches, machine learning techniques and a number of other methods used in bioinformatics for this type of analysis. The use of a statistical approach helped us focus on the most relevant characteristic, allowing us to provide a detailed picture of the molecular mechanisms of resistance.

The results of the structural analysis show that mutations can be clustered according to their mechanism of resistance. For example, mutations causing the loss of negatively charged amino acids, will lead to a change of the charge distribution in the drug binding site which interferes with quinolone binding.

2. Development of Quinores3D web server.

Although bioinformatics databases exist collecting AR genes, point mutations and associated phenotypes (e.g. CARD³⁰), no resources for protein structural information related to AR were made so far available.

The pipeline developed for the structural analysis was incorporated in a publicly available web server (<http://bioinfoibpm.cloud.ba.infn.it/quinores3d/index.html>). In particular Quinores3D Finder is able to perform a structural analysis given an input topoisomerase II protein or nucleotide sequence or even a complete bacterial genome, and identify point mutations associated with drug resistance. Results from the analysis are shown in tables and graphically displayed thanks to viewers embedded in the web server.

We also developed Quinores3D db, a database collecting structural information and features about the mutations related to quinolone resistance.

3. Development of new methods for the identification of point mutations associated with AR.

Several tools for the identification of mutations associated with AR are available, such as RGI³⁰, PointFinder²⁶ or AMRfinder³⁵. Although they are widely used, these tools present some issues: they are limited to a set of bacterial species²⁶; they suffer from high false-susceptible (false-negative) and/or false-resistant

(false-positive) rates¹⁰¹; they can detect known mutations in user sequences but do not allow inference of new mutations potentially associated with antibiotic resistance. To overcome these limitations, we developed a machine learning classifier, which can be used to probably infer new point mutations related to quinolone resistance in different bacterial species. The classifier uses the Random Forest algorithm and achieved an accuracy of 94 % on the test set. At the moment this prediction tool is specific for quinolone antibiotics only.

We are currently working on the application to other classes of antibiotics of the approaches developed in this work for the analysis and prediction of point mutations associated with AR to quinolones.

Our aim is to extend *Quinores3D db*, *Quinores3D Finder*, as well as the classifier to the study, structural analysis, and inference of point mutations associated with resistance to antibiotics other than quinolones. This

will be possible whenever the three-dimensional structure of a drug target exists or can be modelled, so that structural information can be extracted, annotated, made available through a database and used to train, test and validate a machine learning algorithm.

6. Bibliography

1. Clardy, J., Fischbach, M. A. & Currie, C. R. The natural history of antibiotics. *Curr. Biol.* **19**, R437–R441 (2009).
2. Hutchings, M., Truman, A. & Wilkinson, B. Antibiotics: past, present and future. *Curr. Opin. Microbiol.* **51**, 72–80 (2019).
3. Gelpi, A., Gilbertson, A. & Tucker, J. D. Magic bullet: Paul Ehrlich, Salvarsan and the birth of venereology. *Sex. Transm. Infect.* **91**, 68–69 (2015).
4. Johnson, A. Antimicrobial Agents: Antibacterials and Antifungals Andre Bryskier, Ed. ASM Press, Washington, USA, 2005. ISBN 1-55581-237-6. \$189.95, 1456 pp. *J. Antimicrob. Chemother.* **58**, 231 (2006).
5. Kapoor, G., Saigal, S. & Elongavan, A. Action and resistance mechanisms of antibiotics: A guide for clinicians. *J. Anaesthesiol. Clin. Pharmacol.* **33**, 300–305 (2017).
6. Takenouchi, T. *et al.* Hydrophilicity of quinolones is not an exclusive factor for decreased activity in efflux-mediated resistant mutants of *Staphylococcus aureus*. *Antimicrob. Agents Chemother.* **40**, 1835–1842 (1996).
7. Correia, S., Poeta, P., Hébraud, M., Capelo, J. L. & Igrejas, G. Mechanisms of quinolone action and resistance: where do we stand? *J. Med. Microbiol.* **66**, 551–559 (2017).
8. Pham, T. D. M., Ziora, Z. M. & Blaskovich, M. A. T. Quinolone antibiotics. *Medchemcomm* **10**, 1719–1739 (2019).
9. King, D. E., Malone, R. & Lilley, S. H. New classification and update

- on the quinolone antibiotics. *Am. Fam. Physician* **61**, 2741–2748 (2000).
10. Laponogov, I. *et al.* Structure of an ‘open’ clamp type II topoisomerase-DNA complex provides a mechanism for DNA capture and transport. *Nucleic Acids Res.* **41**, 9911–9923 (2013).
 11. Laponogov, I. *et al.* Structural insight into the quinolone-DNA cleavage complex of type IIA topoisomerases. *Nat. Struct. Mol. Biol.* **16**, 667–669 (2009).
 12. Champoux, J. J. DNA Topoisomerases: Structure, Function. *Ann Rev Biochem* **70**, 369–413 (2001).
 13. Petrella, S. *et al.* Overall Structures of Mycobacterium tuberculosis DNA Gyrase Reveal the Role of a Corynebacteriales GyrB-Specific Insert in ATPase Activity. *Structure* **27**, 579-589.e5 (2019).
 14. Weidlich, D. & Klostermeier, D. Functional interactions between gyrase subunits are optimized in a species-specific manner. *J. Biol. Chem.* **295**, 2299–2312 (2020).
 15. Vanden Broeck, A., Lotz, C., Ortiz, J. & Lamour, V. Cryo-EM structure of the complete E. coli DNA gyrase nucleoprotein complex. *Nat. Commun.* **10**, 4935 (2019).
 16. Laponogov, I. *et al.* Exploring the active site of the Streptococcus pneumoniae topoisomerase IV-DNA cleavage complex with novel 7,8-bridged fluoroquinolones. *Open Biol.* **6**, (2016).
 17. Laponogov, I. *et al.* Breakage-Reunion Domain of Streptococcus pneumoniae Topoisomerase IV: Crystal Structure of a Gram-Positive Quinolone Target. *PLoS One* **2**, e301 (2007).
 18. Li, B. & Webster, T. J. Bacteria antibiotic resistance: New challenges and opportunities for implant-associated orthopedic infections. *J.*

- Orthop. Res.* **36**, 22–32 (2018).
19. Wright, G. D. Q&A: Antibiotic resistance: Where does it come from and what can we do about it? *BMC Biol.* **8**, (2010).
 20. Munita, J. M., Arias, C. A., Unit, A. R. & Santiago, A. De. HHS Public Access Mechanisms of Antibiotic Resistance. *HHS Public Access* **4**, 1–37 (2016).
 21. Hooper, D. C. & Jacoby, G. A. Mechanisms of drug resistance: quinolone resistance. *Ann. N. Y. Acad. Sci.* **1354**, 12–31 (2015).
 22. Lupala, C., Gomez-Gutierrez, P. & Perez, J. Molecular Determinants of the Bacterial Resistance to Fluoroquinolones: A Computational Study. *Curr. Comput. Aided-Drug Des.* **9**, 281–288 (2013).
 23. Aldred, K. J., McPherson, S. A., Turnbough, C. L., Kerns, R. J. & Osheroff, N. Topoisomerase IV-quinolone interactions are mediated through a water-metal ion bridge: Mechanistic basis of quinolone resistance. *Nucleic Acids Res.* **41**, 4628–4639 (2013).
 24. Van Camp, P. J., Haslam, D. B. & Porollo, A. Bioinformatics approaches to the understanding of molecular mechanisms in antimicrobial resistance. *Int. J. Mol. Sci.* **21**, (2020).
 25. Boolchandani, M., D’Souza, A. W. & Dantas, G. Sequencing-based methods and resources to study antimicrobial resistance. *Nat. Rev. Genet.* **20**, 356–370 (2019).
 26. Zankari, E. *et al.* PointFinder: A novel web tool for WGS-based detection of antimicrobial resistance associated with chromosomal point mutations in bacterial pathogens. *J. Antimicrob. Chemother.* **72**, 2764–2768 (2017).
 27. Rishishwar, L., Petit 3rd, R. A., Kraft, C. S. & Jordan, I. K. Genome sequence-based discriminator for vancomycin-intermediate

- Staphylococcus aureus. *J. Bacteriol.* **196**, 940–948 (2014).
28. Bradley, P. *et al.* Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. *Nat. Commun.* **6**, 10063 (2015).
 29. Davis, J. J. *et al.* Antimicrobial Resistance Prediction in PATRIC and RAST. *Sci. Rep.* **6**, 27930 (2016).
 30. Alcock, B. P. *et al.* CARD 2020: Antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **48**, D517–D525 (2020).
 31. Zankari, E. *et al.* Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.* **67**, 2640–2644 (2012).
 32. Gupta, S. K. *et al.* ARG-ANNOT, a New Bioinformatic Tool To Discover Antibiotic Resistance Genes in Bacterial Genomes. *Antimicrob. Agents Chemother.* **58**, 212 LP – 220 (2014).
 33. Yang, Y. *et al.* ARGs-OAP: online analysis pipeline for antibiotic resistance genes detection from metagenomic data using an integrated structured ARG-database. *Bioinformatics* **32**, 2346–2351 (2016).
 34. Hunt, M. *et al.* ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microb. genomics* **3**, e000131–e000131 (2017).
 35. Feldgarden, M. *et al.* Validating the AMRFINDER tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates. *Antimicrob. Agents Chemother.* **63**, 1–20 (2019).
 36. Inouye, M. *et al.* SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med.* **6**, 90 (2014).
 37. Rowe, W. *et al.* Search Engine for Antimicrobial Resistance: A Cloud

- Compatible Pipeline and Web Interface for Rapidly Detecting Antimicrobial Resistance Genes Directly from Sequence Data. *PLoS One* **10**, e0133492 (2015).
38. Kaminski, J. *et al.* High-Specificity Targeted Functional Profiling in Microbial Communities with ShortBRED. *PLOS Comput. Biol.* **11**, e1004557 (2015).
 39. de Man, T. J. B. & Limbago, B. M. SSTAR, a Stand-Alone Easy-To-Use Antimicrobial Resistance Gene Predictor. *mSphere* **1**, e00050-15 (2016).
 40. Clausen, P. T. L. C., Zankari, E., Aarestrup, F. M. & Lund, O. Benchmarking of methods for identification of antimicrobial resistance genes in bacterial whole genome data. *J. Antimicrob. Chemother.* **71**, 2484–2488 (2016).
 41. Berglund, F. *et al.* Identification and reconstruction of novel antibiotic resistance genes from metagenomes. *Microbiome* **7**, 52 (2019).
 42. Arango-Argoty, G. *et al.* DeepARG: A deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome* **6**, 1–15 (2018).
 43. McArthur, A. G. *et al.* The comprehensive antibiotic resistance database. *Antimicrob. Agents Chemother.* **57**, 3348–3357 (2013).
 44. Shanthirabalan, S., Chomilier, J. & Carpentier, M. Structural effects of point mutations in proteins. *Proteins Struct. Funct. Bioinforma.* **86**, 853–867 (2018).
 45. Feyfant, E., Sali, A. & Fiser, A. Modeling mutations in protein structures. *Protein Sci.* **16**, 2030–2041 (2007).
 46. Peng, Y., Alexov, E. & Basu, S. Structural Perspective on Revealing and Altering Molecular Functions of Genetic Variants Linked with

- Diseases. *Int. J. Mol. Sci.* **20**, 548 (2019).
47. Hurst, J. M. *et al.* The SAAPdb web resource: A large-scale structural analysis of mutant proteins. *Hum. Mutat.* **30**, 616–624 (2009).
 48. Pandurangan, A. P., Ascher, D. B., Thomas, S. E. & Blundell, T. L. Genomes, structural biology and drug discovery: combating the impacts of mutations in genetic disease and antibiotic resistance. *Biochem. Soc. Trans.* **45**, 303–311 (2017).
 49. Teng, S., Srivastava, A. K., Schwartz, C. E., Alexov, E. & Wang, L. Structural assessment of the effects of amino acid substitutions on protein stability and protein protein interaction. *Int. J. Comput. Biol. Drug Des.* **3**, 334–349 (2010).
 50. Engholm, D. H., Kilian, M., Goodsell, D. S., Andersen, E. S. & Kjærgaard, R. S. A visual review of the human pathogen *Streptococcus pneumoniae*. *FEMS Microbiol. Rev.* **41**, 854–879 (2017).
 51. Weiser, J. N., Ferreira, D. M. & Paton, J. C. *Streptococcus pneumoniae*: transmission, colonization and invasion. *Nat. Rev. Microbiol.* **16**, 355–367 (2018).
 52. Hoffer, E. Machine Learning을 이용한 자동 돌발상황검지. *Seoul Stud.* **6**, 71–80 (2005).
 53. Bhaskar, H., Hoyle, D. C. & Singh, S. Machine learning in bioinformatics : A brief survey and recommendations for practitioners. **36**, 1104–1125 (2006).
 54. Larrañaga, P. *et al.* Machine learning in bioinformatics. *Brief. Bioinform.* **7**, 86–112 (2006).
 55. Li, Y. *et al.* Deep learning in bioinformatics: Introduction, application,

- and perspective in the big data era. *Methods* **166**, 4–21 (2019).
56. Oyelade, J. *et al.* Clustering Algorithms: Their Application to Gene Expression Data. *Bioinform. Biol. Insights* **10**, 237–253 (2016).
 57. Couronné, R., Probst, P. & Boulesteix, A. L. Random forest versus logistic regression: A large-scale benchmark experiment. *BMC Bioinformatics* **19**, 1–14 (2018).
 58. Chen, K. & Kurgan, L. A. Neural Networks in Bioinformatics BT - Handbook of Natural Computing. in (eds. Rozenberg, G., Bäck, T. & Kok, J. N.) 565–583 (Springer Berlin Heidelberg, 2012).
doi:10.1007/978-3-540-92910-9_18.
 59. Varoquaux, G. *et al.* Scikit-learn. *GetMobile Mob. Comput. Commun.* **19**, 29–33 (2015).
 60. Chen, X. & Ishwaran, H. Random forests for genomic data analysis. *Genomics* **99**, 323–329 (2012).
 61. Chen, Y.-Y., Lin, Y.-H., Kung, C.-C., Chung, M.-H. & Yen, I.-H. Design and Implementation of Cloud Analytics-Assisted Smart Power Meters Considering Advanced Artificial Intelligence as Edge Analytics in Demand-Side Management for Smart Homes. *Sensors (Basel)*. **19**, 2047 (2019).
 62. Jiao, Y. & Du, P. Performance measures in evaluating machine learning based bioinformatics predictors for classifications. *Quantitative Biology* vol. 4 320–330 (2016).
 63. Coordinators, N. R. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **44**, D7–D19 (2016).
 64. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
 65. Edgar, R. C. MUSCLE: multiple sequence alignment with high

- accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
66. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
67. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).
68. Karavirta, V. & Shaffer, C. A. JSAV: The JavaScript algorithm visualization library. in *Annual Conference on Innovation and Technology in Computer Science Education, ITiCSE* 159–164 (2013). doi:10.1145/2462476.2462487.
69. Eswar, N. *et al.* Comparative protein structure modeling using Modeller. *Curr. Protoc. Bioinforma.* **Chapter 5**, Unit-5.6 (2006).
70. Söding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33**, W244–W248 (2005).
71. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
72. Corbett, K. D., Shultzaberger, R. K. & Berger, J. M. The C-terminal domain of DNA gyrase A adopts a DNA-bending β -pinwheel fold. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 7293 LP – 7298 (2004).
73. Agrawal, A. *et al.* Mycobacterium tuberculosis DNA gyrase ATPase domain structures suggest a dissociative mechanism that explains how ATP hydrolysis is coupled to domain motion. *Biochem. J.* **456**, 263–273 (2013).
74. Veselkov, D. A. *et al.* Structure of a quinolone-stabilized cleavage

- complex of topoisomerase IV from *Klebsiella pneumoniae* and comparison with a related *Streptococcus pneumoniae* complex. *Acta Crystallogr. Sect. D Struct. Biol.* **72**, 488–496 (2016).
75. Hsieh, T. J., Farh, L., Huang, W. M. & Chan, N. L. Structure of the topoisomerase IV C-terminal domain: A broken β -propeller implies a role as geometry facilitator in catalysis. *J. Biol. Chem.* **279**, 55587–55593 (2004).
76. Laponogov, I. *et al.* Trapping of the transport-segment DNA by the ATPase domains of a type II topoisomerase. *Nat. Commun.* **9**, 2579 (2018).
77. Shen, M.-Y. & Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **15**, 2507–2524 (2006).
78. Benkert, P., Biasini, M. & Schwede, T. Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics* **27**, 343–350 (2011).
79. Guex, N. & Peitsch, M. C. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **18**, 2714–2723 (1997).
80. Jurrus, E. *et al.* Improvements to the APBS biomolecular solvation software suite. *Protein Sci.* **27**, 112–128 (2018).
81. Pettersen, E. F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
82. Untergasser, A. *et al.* Primer3--new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115–e115 (2012).
83. Satopää, V., Albrecht, J., Irwin, D. & Raghavan, B. Finding a ‘kneedle’ in a haystack: Detecting knee points in system behavior.

- Proc. - Int. Conf. Distrib. Comput. Syst.* 166–171 (2011)
doi:10.1109/ICDCSW.2011.20.
84. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
 85. Capra, J. A. & Singh, M. Predicting functionally important residues from sequence conservation. *Bioinformatics* **23**, 1875–1882 (2007).
 86. Johansson, F. & Toh, H. A comparative study of conservation and variation scores. *BMC Bioinformatics* **11**, 388 (2010).
 87. Adami, C. Information theory in molecular biology. *Physics of Life Reviews* vol. 1 3–22 (2004).
 88. Sen, S., Dey, A., Chowdhury, S., Maulik, U. & Chattopadhyay, K. Understanding the evolutionary trend of intrinsically structural disorders in cancer relevant proteins as probed by Shannon entropy scoring and structure network analysis. *BMC Bioinformatics* **19**, (2019).
 89. Bisong, E. Google Colaboratory BT - Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners. in (ed. Bisong, E.) 59–64 (Apress, 2019).
doi:10.1007/978-1-4842-4470-8_7.
 90. Kawashima, S. *et al.* AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.* **36**, D202-5 (2008).
 91. Chen, Z. *et al.* iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* **34**, 2499–2502 (2018).
 92. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

93. Madurga, S., Sánchez-Céspedes, J., Belda, I., Vila, J. & Giralt, E. Mechanism of binding of fluoroquinolones to the quinolone resistance-determining region of DNA gyrase: Towards an understanding of the molecular basis of quinolone resistance. *ChemBioChem* **9**, 2081–2086 (2008).
94. Murray, G. L. *et al.* Increasing Macrolide and Fluoroquinolone Resistance in *Mycoplasma genitalium*. *Emerg. Infect. Dis.* **23**, 809–812 (2017).
95. Zhang, G., Tian, W., Wang, C. & Feng, J. Identification of a novel resistance mutation in *parE* that confers high-level resistance to moxifloxacin in *Streptococcus pneumoniae*. *J. Antimicrob. Chemother.* **67**, 2773–2774 (2012).
96. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132 (1982).
97. Velema, W. A. *et al.* Ciprofloxacin-Photoswitch Conjugates: A Facile Strategy for Photopharmacology. *Bioconjug. Chem.* **26**, 2592–2597 (2015).
98. Yoshida, H., Bogaki, M., Nakamura, M., Yamanaka, L. M. & Nakamura, S. Quinolone resistance-determining region in the DNA gyrase *gyrB* gene of *Escherichia coli*. *Antimicrob. Agents Chemother.* **35**, 1647–1650 (1991).
99. Neill, J. O. ' . Antimicrobial Resistance: Tackling a crisis for the health and wealth of nations The Review on Antimicrobial Resistance Chaired. (2014).
100. Ventola, C. L. The Anti - biotic Resistance Crisis Part 2 : Management Strategies and New Agents. **40**, 344–352 (2015).
101. Mahfouz, N., Ferreira, I., Beisken, S., von Haeseler, A. & Posch, A. E.

Large-scale assessment of antimicrobial resistance marker databases for genetic phenotype prediction: a systematic review. *J. Antimicrob. Chemother.* 1–10 (2020) doi:10.1093/jac/dkaa257.