Contents

“Owing to this struggle for life, any variation, however slight and from whatever cause proceeding, if it be in any degree profitable to an individual of any species, in its infinitely complex relations to other organic beings and to external nature, will tend to the preservation of that individual, and will generally be inherited by its offspring. (...) I have called this principle, by which each slight variation, if useful, is preserved, by the term of Natural Selection, in order to mark its relation to man’s power of selection.”

Charles Darwin, On the Origin of Species, 1859

1

Drug resistance-associated mutations

In this chapter we present our work on identifying drug resistance-associated mutations based on comparative analysis of whole-genome sequences of closely related bacterial strains. In particular, we present GWAMAR, the tool we have developed to support this type of analysis. In section 1.1, we describe the idea behind our approach and review some related work. We also introduce the basic concepts and notations. In section 1.2, we describe in detail the methodology of GWAMAR. Notably, it uses eCAMBer, described in chapter **??**, for identification of genetic variations (mutations) among the set of considered strains, which constitute the genotype data. As a part of this section, we also present *weighted support* (WS) and *tree-generalized hypergeometric* (TGH) score — two statistics we propose for identifying of drug resistance associations. Additionally, we propose a Rank-based metascore (RBM) for combining multiple scores into one in order to compromise between different approaches used to define different scores. In section 1.3, we present and discuss results obtained by applying GWAMAR to three datasets — one for *S. aureus* and two for *M. tuberculosis*. The presented results show that GWAMAR can be successfully used for identification of drug resistance-associated mutations.

1.1 Introduction

Genome-Wide Association Studies (GWAS) have been successfully applied to associate human mutations with phenotype of various human diseases and traits (Manolio, 2010; Stadler et al., 2010; Davies et al., 2011).

The recent progress in genome-sequencing technologies, continuously decreasing the cost of sequencing of bacterial genomes (Loman et al., 2012), enables the use of similar approaches for genotype-phenotype mapping in bacteria.

The potential of the use of whole-genome comparative approaches to study drug resistance and host-pathogen interactions in bacteria has been recently proposed (Khor and Hibberd, 2012; Read and Massey, 2014).

1.1.1 Genotype data

The input genotype data for these studies usually comes from in-house sequencing, rather than publicly available data. This might be caused by the problematic use of the publicly available data. First, as we noticed in the previous chapter, the inconsistent and poor-quality annotations of publicly available strains may complicate that analysis. Second, the phenotype data with respect to drug susceptibility tests are spread throughout the literature and are not easy to collect.

In the previous chapter of this work, we presented CAMBer and eCAMBer — the tools to support comparative analysis of multiple bacterial strains — thus addressing the first issue. In order to overcome the second issue, we have to perform a careful search of the literature for results of drug susceptibility tests of the strains considered.

1.1.2 Phenotype data

Minimum Inhibitory Concentration (MIC) is the most commonly used measure to quantify drug resistance in bacteria. It is the lowest concentration of an antibiotic which inhibits visible growth of a colony of bacteria after overnight incubation. The detailed guidelines for the procedure of drug susceptibility testing are published by bodies such as Clinical and Laboratory Standards Institute (CLSI), British Society for Antimicrobial Chemotherapy (BSAC), and The European Committee on Antimicrobial Susceptibility Testing (EUCAST). The guidelines also contain information on MIC breakpoints to assign drug resistance or drug susceptibility. Sometimes also the third class of intermediate resistance is distinguished.

Most of the sources reporting results of drug susceptibility testing provide only information on the outcome status, rather than particular MIC values. Thus, in our study we use only three classes of drug resistance: drug susceptible, intermediate drug resistant and drug resistant.

For the purpose of this work, we have collected the phenotype data for drug resistance from the following sources: (i) publications issued together with the fully sequenced genomes; (ii) NARSA project (`http://www.narsa.net`); (iii) email exchange with the authors of some publications; and (iv) other publications found by searching of the related literature.

1.1.3 Gold standard associations

One problem we faced during the project was caused by the relatively small number of positive associations available in the databases, which would constitute the gold standard data to assess the accuracy of our method.

Nevertheless, there are known genes and point mutations responsible for some of the drug resistance mechanisms. However, these are spread over various studies and are therefore not easy to gather.

One attempt to collect the information on genetic changes associated with drug resistance into a database is the Antibiotic Drug Resistance Database (ARDB) developed by Liu and Pop (2009). However, this database focuses on genes associated with drug resistance rather than particular point mutations within them. We use data available in this database as our gold standard for the case study on the *S. aureus* dataset, presented in the results section of the chapter.

Another species-specific database of drug resistance-associated mutations in *M. tuberculosis* is the Tuberculosis Drug Resistance Mutation Database (TB-DReaMDB) developed by Sandgren et al. (2009). This database provides detailed information on a set of 1230 associations between drugs and point mutations. Furthermore, it distinguishes a subset of *high-confidence* mutations which were often reported in the literature. We use data available in this database as our gold standard for the case study on the two *M. tuberculosis* datasets, presented in the results section of the chapter.

1.1.4 Phylogenetic information

In this work we investigate the potential of the use of phylogenetic information in identifying drug resistance-associated mutations. In particular, we propose two association scores, called TGH and WS, based on the phylogenetic information.

The rationale for our approach is based on two known phenomena. First, the bacteria isolated from close-distance locations of each other tend to have similar genome sequences. As a result, subtrees of the phylogenetic trees tend to correspond to geographic locations (Daubin et al., 2003).

Second, although the phenomenon of genomic convergence is unlikely in general, it is rather common in case of mutations which are subject to evolutionary pressure caused by drug treatment (Hazbón et al., 2008; Farhat et al., 2013). Thus, drug resistance-associated mutations tend to be independent of geographic location and therefore more widely distributed over the tree, as opposed to mutations driven by other environmental factors which tend to concentrate in small subtrees.

Hence, mutations predicted to occur independently multiple times in the evolutionary history of the bacterial strains are more likely to be associated with drug resistance, rather than with other environmental factors (Hazbón et al., 2008). A conceptually similar approach has been taken by Dutheil (2012) to identify co-evolving mutations in protein sequences.

We note however, it is only an approximation to represent the evolutionary history of bacteria as a tree. It has been debated that, in the presence of HGT mechanisms in bacteria, their evolutionary history may be better represented as a network rather than a tree (Philippe and Douady, 2003). On the other hand, some estimations show that the effect of HGT on the overall evolution is limited and does not preclude the use of phylogenetic trees (Daubin et al., 2003; Boto, 2010). We leave the possibility of using other representations of the evolutionary history of bacteria as a subject of further research.

1.1.5 Basic definitions

In this work, we consider a set $\mathcal{S}$ of closely related bacterial genomes. Typically, this is a set of strains within the same species of bacteria.

Then, we represent the available drug resistance information as a set of *drug resistance profiles* $\mathcal{R}$, where each drug resistance profile $r \in \mathcal{R}$ is represented as a vector:

$$r : \mathcal{S} \rightarrow \{\text{'S'}, \text{'I'}, \text{'R'}, \text{'?'}\}. \tag{1.1}$$

Here, 'S', 'I', 'R' denote that a given strain is known to be drug susceptible, intermediate-resistant, or resistant, respectively. We indicate, using question mark '?', that the drug resistance status of a strain is unknown. We call a drug resistance profile *complete* if it does not contain question marks.

The genotype data consists of a set of genetic mutations of three types:

- point mutations (in amino-acid sequences),
- gene gain/losses,
- promoter mutations.

In our approach we exclude synonymous SNPs as, according to our knowledge, there are no known examples of synonymous mutations associated with drug resistance.

Each mutation is represented as a piece of information adequate for the type of the mutation (such as gene identifier of the corresponding gene family) and a vector called *mutation profile*:

$$v : \mathcal{S} \rightarrow \Sigma. \tag{1.2}$$

Here, for each point mutation, we keep the information on its position in the multiple alignment of its corresponding gene family and the information on the gene family identifier. The mutation profile for each point mutation is determined based on its corresponding column in the multiple alignment. In that case $\Sigma = \Sigma_{AA}$ denotes the set of twenty amino acids. We also assume Σ_{AA} contains the '-', symbol for the gap in the corresponding multiple alignment and the '?' symbol if the gene sequence is unknown for a given strain. We take into account only columns which contain at least two different characters (ignoring '?').

Next, for each gene gain/loss, we keep the information on its corresponding gene family identifier. For such a mutation, its mutation profile is determined based on the presence or absence of a gene in the corresponding gene family for a given strain. Thus, $\Sigma = \{\text{'L'}, \text{'G'}\}$, where $v(S) = \text{'L'}$ means that the gene is absent in strain S, whereas $v(S) = \text{'G'}$ means that the gene is present in strain S.

Finally, for each promoter mutation, we keep the information on its position in the multiple alignment of promoter sequences for the corresponding gene family and the information on the gene family identifier. The mutation profile for each promoter mutation is determined based on its corresponding column in the multiple alignment. In that case $\Sigma = \Sigma_{NT}$ denotes the set of four different nucleotides together with the '-' symbol for gaps in the corresponding multiple alignment and the '?' symbol if the gene promoter sequence is unknown for a given strain.

Analogously, we call a mutation profile *complete* if it does not contain question marks.

It should be noted that potentially multiple mutations (for example point mutations at different positions in the genome) may have identical mutation profiles. In that situation the mutations would essentially carry the same information about their mutation profiles. Thus, we also introduce an auxiliary concept called *binary mutation profile*. Let $S^* \in \mathcal{S}$ denote the reference strain and $S \in \mathcal{S}$ denote any strain. Then, for a given *mutation profile* v, its corresponding binary mutation profile

$$b_v : \mathcal{S} \to \{\text{'0'}, \text{'1'}, \text{'?'}\}, \tag{1.3}$$

is defined as follows:

$$b_v(S) = \begin{cases} \text{'?'} & \text{if } v(S) = \text{'?'} \\ \text{'0'} & \text{if } v(S) = v(S^*) \\ \text{'1'} & \text{otherwise} \end{cases} \tag{1.4}$$

Analogous to mutation profiles, we call a binary mutation profile *complete* if it does not contain question marks.

We say that a genetic change (mutation) m is *present* in strain $S \in \mathcal{S}$ if for its corresponding mutation profile v, $b_v(S) = \text{'1'}$; otherwise we say that the mutation m is *absent* in strain S.

1.1.6 Problem setting

Finally, we define the problem which we address here: given a list of mutations and a list of drug resistance profiles, produce an ordered list of associations between the phenotype and genotype data (represented as drug resistance and mutation profiles) such that the top-scored associations are the most likely to be real.

1.2 GWAMAR: Genome-wide assessment of mutations associated with drug resistance in bacteria

In this section, we present details of GWAMAR, the tool we have developed for genome-wide assessment of mutations associated with drug resistance. The presentation includes the preprocessing of input data; computation of the association scores and results obtained by applying the tool to datasets for *M. tuberculosis* and *S. aureus*.

1.2.1 The pipeline of GWAMAR

GWAMAR is designed as a pipeline. It first employs eCAMBer, the tool described in the previous chapter, to perform three preliminary steps: (i) downloading of genome sequences and annotations for the set of multiple bacterial strains in question, (ii) consolidation of the genome annotations, (iii) identification of homologous gene families; see Figure 1.1.

In the next step eCAMBer identifies the set of genetic variations and represents them as mutations profiles. As described in section 1.1.5, three types of mutations are considered: (i) point mutations in amino-acid sequences, (ii) point mutations in promoter regions (-50bp downstream the corresponding TIS), (iii) gene gain/losses.

Here, each gene gain/loss mutation profile is determined based on the presence/absence of elements of the corresponding gene family among the strains.

For each identified gene family, eCAMBer employs MUSCLE (Edgar, 2004), to compute its multiple sequence alignment for the set of corresponding amino-acid sequences. Similarly, it uses MUSCLE to compute a multiple sequence alignment for the set of corresponding promoter sequences.

Next, eCAMBer transforms each column in the computed multiple alignment into a mutation profile, as long as at least one character in that column differs (there is a mutation present); see Figure 1.1.

Also, eCAMBer supports use of PHYLIP (Felsenstein, 2005) and PhyML (Guindon et al., 2010) — the software for reconstruction of the phylogenetic tree based on the maximal-likelihood approach.

In the next step, for the selected reference strain, GWAMAR computes binary mutation profiles for each mutation profile, based on formula 1.4. Since multiple mutation profiles may correspond to a binary mutation profile, this step significantly reduces the number of pairs of profiles (resistance and mutation profiles) to be scored.

Finally, GWAMAR computes several statistical scores to associate drug resistance profiles to the mutation profiles, including mutual information (MI), odds ratio (OR), hypergeometric (H) score, weighted support (WS), and the tree-generalized hypergeometric (TGH) score. Additionally, it implements a score we called Rank-based metascore (RBM) which for combining multiple scores into one in order to compensate for weaknesses of different individual scores.

Figure 1.1 illustrates the overall data-processing flow implemented in GWAMAR.

1.2.2 Association scores

Here we present the association scores implemented in GWAMAR to score pairs of binary mutation and drug-resistance profiles. These scores include statistics commonly used in various associations studies, such as mutual information (Wu et al., 2012), odds ratio (Clarke et al., 2011), hypergeometric test (Cabrera et al., 2012). It also computes weighted support and tree-generalized hypergeometric score — the newly proposed statistics to incorporate the phylogenetic information. Moreover, it implements the Rank-based metascore for combining multiple scores into one.

For a given binary mutation profile $b\mathcal{B}$ and a given drug resistance profile $r\mathcal{R}$, we introduce the following auxiliary notations:

- $\mathcal{S}_1^R = \{S \in \mathcal{S} : b(S) = \text{'1'} \wedge r(S) = \text{'R'}\}$,
- $\mathcal{S}_0^R = \{S \in \mathcal{S} : b(S) = \text{'0'} \wedge r(S) = \text{'R'}\}$,

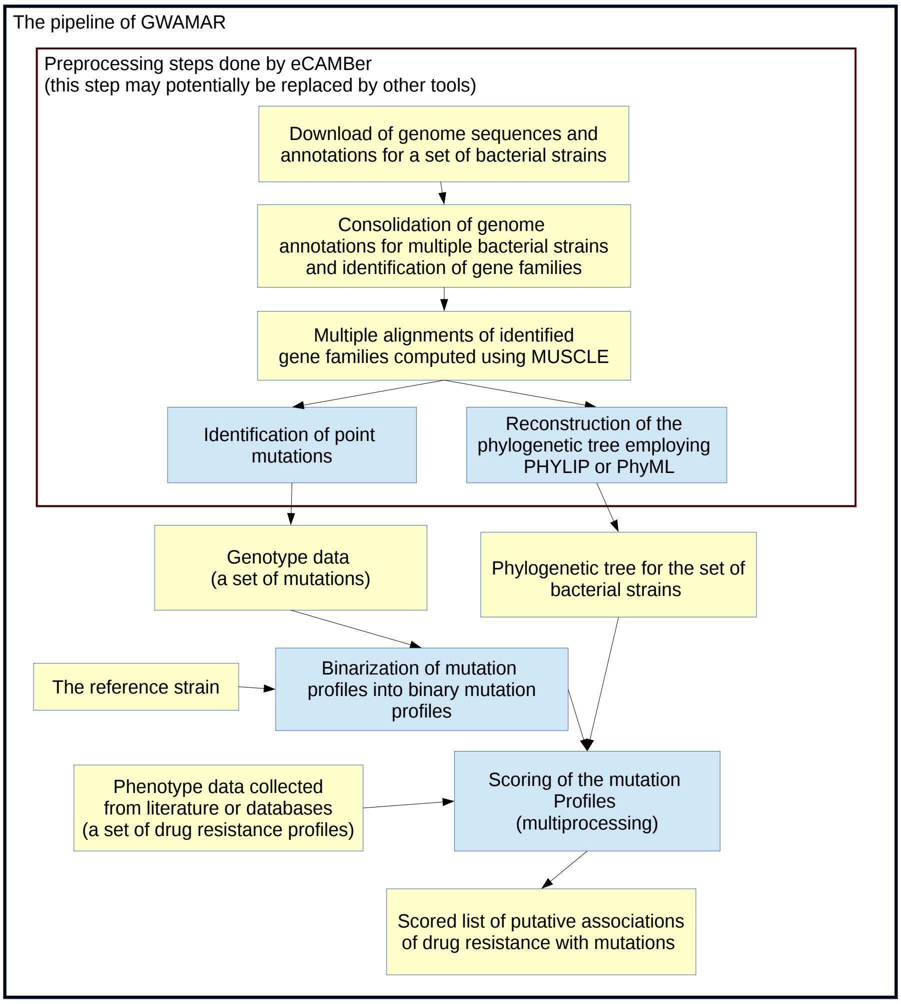

The pipeline of GWAMAR
Preprocessing steps done by eCAMBer
(this step may potentially be replaced by other tools)
Download of genome sequences and annotations for a set of bacterial strains
Consolidation of genome annotations for multiple bacterial strains and identification of gene families
Multiple alignments of identified gene families computed using MUSCLE
Identification of point mutations
Reconstruction of the phylogenetic tree employing PHYLIP or PhyML
Genotype data (a set of mutations)
Phylogenetic tree for the set of bacterial strains
The reference strain
Binarization of mutation profiles into binary mutation profiles
Phenotype data collected from literature or databases (a set of drug resistance profiles)
Scoring of the mutation Profiles (multiprocessing)
Scored list of putative associations of drug resistance with mutations

Figure 1.1: Schema of the pipeline of GWAMAR. For a set of considered bacterial strains, the input data for GWAMAR consists of (i) a set of mutations; (ii) a set of drug resistance profiles; and (iii) optional, phylogenetic tree for the set of bacterial strains. Typically the set of mutation profiles is generated using eCAMBer, which is able to download the genome sequences and annotations for the set of bacterial strains, identify point mutations based on multiple alignments, and reconstruct the phylogenetic tree of the considered bacterial strains. Assuming the genotype data is preprocessed, the first step of GWAMAR is to compute binary mutation profiles for all the mutations. This step significantly reduces the number of profiles considered. Finally, GWAMAR implements several statistical scores to associate drug resistance profiles with mutation profiles. These include: mutual information, odds ratio, hypergeometric score, weighted support, tree-generalized hypergeometric and the Rank-based metascore. As a result, we obtain ordered lists of drug resistance associations, where the top-scored associations are the most likely to be real.

- $\mathcal{S}_1^I = \{S \in \mathcal{S} : b(S) = \text{'1'} \wedge r(S) = \text{'I'}\}$,

- $\mathcal{S}_0^I = \{S \in \mathcal{S} : b(S) = \text{'0'} \wedge r(S) = \text{'I'}\}$,

- $\mathcal{S}_1^S = \{S \in \mathcal{S} : b(S) = \text{'1'} \wedge r(S) = \text{'S'}\}$,

- $\mathcal{S}_0^S = \{S \in \mathcal{S} : b(S) = \text{'0'} \wedge r(S) = \text{'S'}\}$,

- $\mathcal{S}^S = \{S \in \mathcal{S} : r(S) = \text{'S'}\}$,

- $\mathcal{S}^R = \{S \in \mathcal{S} : r(S) = \text{'R'}\}$.

- $\mathcal{S}^I = \{S \in \mathcal{S} : r(S) = \text{'I'}\}$.

- $\mathcal{S}_0 = \{S \in \mathcal{S} : b(S) = \text{'0'}\}$,

- $\mathcal{S}_1 = \{S \in \mathcal{S} : b(S) = \text{'1'}\}$.

Note that, instead of using mutation profiles, we use binary mutation profiles.

1.2.2.1 Odds ratio

For a given binary mutation profile b and drug resistance profile r, we calculate *odds ratio* (OR) score using the following formula:

$$\text{OR}(b, r) = \frac{|\mathcal{S}_1^R| \cdot |\mathcal{S}_0^S|}{max(1, |\mathcal{S}_0^R|) \cdot max(1, |\mathcal{S}_1^S|)} \tag{1.5}$$

Here, we use the *max* function in the denominator to ensure there is no problem with divisibility by 0.

1.2.2.2 Mutual information

For a given binary mutation profile b and a given drug resistance profile r, we calculate *mutual information* (MI) score using the following formula:

$$\text{MI}(b, r) = \sum_{x \in \{\text{'0'},\text{'1'}\}} \sum_{y \in \{\text{'S'},\text{'I'},\text{'R'}\}} \frac{|\mathcal{S}_x^y|}{|\mathcal{S}|} \cdot log(\frac{|\mathcal{S}_x^y| \cdot |\mathcal{S}|}{|\mathcal{S}_x| \cdot |\mathcal{S}^y|}) \tag{1.6}$$

1.2.2.3 Hypergeometric score

For a given binary mutation profile b and a given drug resistance profile r, we calculate hypergeometric (H) score using the following formula:

$$\mathrm{H}(b,r) = -log\big(\sum_{i=|\mathcal{S}^R|}^{|\mathcal{S}|} H(|\mathcal{S}|, |\mathcal{S}^R|, |\mathcal{S}_1|, i)\big) \tag{1.7}$$

where:

$$H(N,K,n,k) = \frac{\binom{K}{k} \cdot \binom{N-K}{n-k}}{\binom{N}{n}} \tag{1.8}$$

Here, we define the hypergeometric score as a minus logarithm of the value typically used in the definition of the hypergeometric test. We use this approach in order to have consistent property for all considered scoring methods, such that the higher the score the more likely drug resistance profile is associated with binary mutation profile.

1.2.2.4 Support

For a given binary mutation b and a given drug resistance profile r, we define *support* (S) as the number of drug-resistant strains with the mutation present minus the number of drug-susceptible strains with the mutation present:

$$\mathrm{S}(b,r) = |\mathcal{S}_1^R| - \alpha(r)|\mathcal{S}_1^S|, \tag{1.9}$$

where:

$$\alpha(r) = \frac{|\mathcal{S}^R|}{|\mathcal{S}^S|} \tag{1.10}$$

Here $\alpha(r)$ is a weight which we use to punish mutations for their presence in drug-susceptible strains. It is defined as the proportion of the number of drug-resistant to the number of drug-susceptible strains, so that occurrences of a mutation are given equal emphasis in drug-resistant and drug-susceptible strains.

1.2.2.5 Weighted support

Although the support is a simple and intuitive score, it does not incorporate any phylogenetic information. For example, let us assume there are two point mutations with the same support 3, where the first mutation covers only drug-resistant strains within one subtree of the phylogenetic tree, whereas the second mutation covers the same number of strains but spread throughout the whole tree. The first mutation is likely to be associated with the phylogeny, driven by some environmental changes. This suggests that the second mutation should have a greater score as it has to be acquired a few times independently during the evolution process.

We propose weighted support (WS) as a score to account for the above situation. For a given phylogenetic tree T, drug resistance profile b, and binary mutation profile b, WS is defined as follows:

$$\mathrm{WS}_T(b, r) = \sum_{S \in \mathcal{S}} w_T(b, r, S)[b(S) = \text{'1'}] \tag{1.11}$$

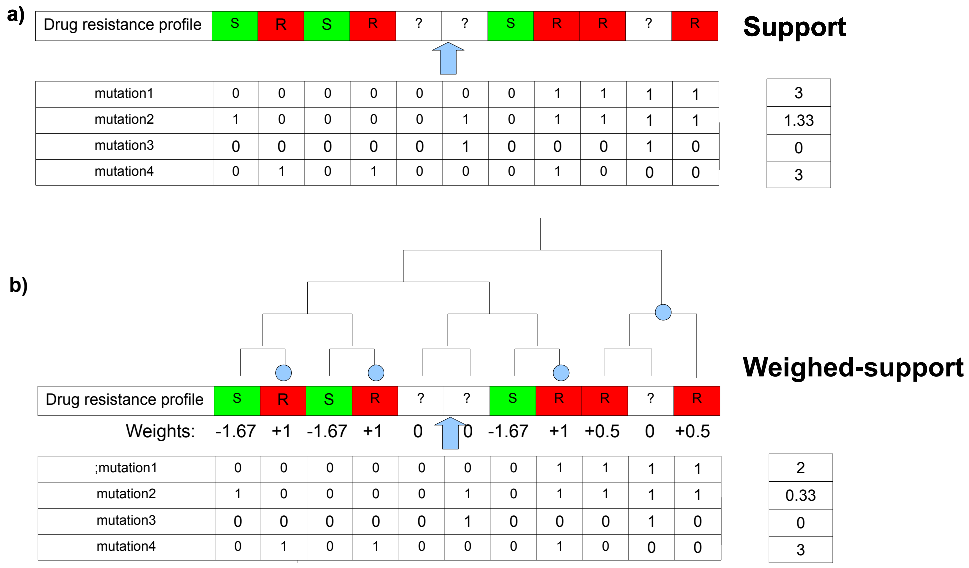
where $w_T(b, r, S)$ is a weight assigned to each cell in a given drug resistance profile.

The weights are assigned in the following way: all drug-susceptible strains are assigned weight $-\alpha(r)$ (defined as above); each drug-resistant strain S is assigned a weight $\frac{1}{n}$, where n is the number of drug-resistant strains in the subtree (containing strain S) determined by its highest parental node, such that the subtree does not contain any drug-susceptible strain in its leaves. All strains without drug resistance information are assigned weights 0.

Note that the support score can also be expressed as weighted support, where $w_T(b, r, S)$ are assigned as $-\alpha(r)$, 1, 0 for drug-susceptible, drug-resistant and strains without drug resistance information, respectively.

Figure 1.2 illustrates the concept of support and weighted support.

In order to make the support scores more comparable between drugs, we introduce normalized versions of the scores, *normalized support* and *normalized weighted support* which denote the respective support value divided by the maximal possible support or weighted support, respectively.



Drug resistance profile	S	R	S	R	?	?	S	R	R	?	R	
mutation1	0	0	0	0	0	0	0	1	1	1	1	3
mutation2	1	0	0	0	0	1	0	1	1	1	1	1.33
mutation3	0	0	0	0	0	1	0	0	0	1	0	0
mutation4	0	1	0	1	0	0	0	1	0	0	0	3

Drug resistance profile	S	R	S	R	?	?	S	R	R	?	R	
Weights:	-1.67	+1	-1.67	+1	0	0	-1.67	+1	+0.5	0	+0.5	
;mutation1	0	0	0	0	0	0	0	1	1	1	1	2
mutation2	1	0	0	0	0	1	0	1	1	1	1	0.33
mutation3	0	0	0	0	0	1	0	0	0	1	0	0
mutation4	0	1	0	1	0	0	0	1	0	0	0	3

Figure 1.2: A schematic example of several mutation profiles and computation of their supports. Light blue circles mark nodes which appear in the definition of weighted support. These are nodes the highest parental nodes (for the leaf nodes corresponding to drug-resistant strains), that their subtrees do not contain any drug-susceptible strains in leaves. The scores (a) support and (b) weighted support are assigned to these mutations. For this drug-resistance profile, the ratio $\alpha(r)$ equals $\frac{5}{3}$.

Statistical significance for WS In order to assess statistical significance of the associations we calculate their *p-values*.

More precisely, for a given drug resistance profile v, let X be the random variable giving support of a random mutation. Then, for a given observed mutation with $Support = c$, its p-value is defined by the following formula:

$$\mathbb{P}(X \geq c) = \sum_{n=1}^{|\mathcal{S}|} \mathbb{P}(X \geq c | N = n) \cdot \mathbb{P}(N = n) \tag{1.12}$$

Here, N is a random variable which denotes the number of mutated strains in a random mutation. For each n the probability $\mathbb{P}(N = n)$ of observing a mutation present in n strains is estimated (as the number of mutations present in n strains to the total number of considered mutations) from the data for point mutation and gene gain/loss profiles separately. The details follow. Assume that weights, for a given drug resistance profile v, take k different values: $l_1, l_2, \ldots, l_k$. For $1 \leq j \leq k$, let m_j be the number of strains which take value l_j. Clearly we have $m_1 + m_2 + \ldots + m_k = |\mathcal{S}|$. Then, the probability $\mathbb{P}(X \geq c | N = n)$ (from the equation 1.12) is given by the formula:

$$\sum_{\substack{0\le n_1\le m_1\\ 0\le n_2\le m_2\\ \dots\\ 0\le n_k\le m_k\\ n_1+n_2+\dots+n_k=n}} \frac{\prod_{j=1}^{k}\binom{m_j}{n_j}}{\binom{|\mathcal{S}|}{n}}\left[\sum_{j=1}^{k} n_j \cdot l_j \ge c\right] \tag{1.13}$$

Here we describe our algorithm for calculating the p-value. It should be clear that the problem reduces to computing $\mathbb{P}(X \ge c|N = n) = \frac{t_c(n)}{\binom{|\mathcal{S}|}{n}}$ for each $0 \le n \le |\mathcal{S}|$, where $t_c(n)$ denotes the number of ways for distributing n ones over $|\mathcal{S}|$ strains, such that the corresponding sum of weights is greater or equal than c. The term $\binom{|\mathcal{S}|}{n}$ is the total number of possible ways for distributing n ones over $|\mathcal{S}|$ strains. Thus, the problem reduces to calculating $t_c(n)$ for each $0 \le n \le |\mathcal{S}|$. Additionally, without any loss of generality, we may assume that the weight levels are strictly decreasing: $l_1 > l_2 > \ldots > l_k$, where $l_k < 0$ and $l_{k-1} \ge 0$.

The algorithm iteratively generates partial combinations (without n_k) starting from the partial combination $(n_1 = m_1, \ldots, n_{k-1} = m_{k-1})$ in the following manner: if j is the highest index of the non-zero n_i in the current partial combination, the next partial combination will be $(n_1, \ldots, n_j - 1, n_{j+1} = m_{j+1}, \ldots, n_{k-1} = m_{k-1})$. The algorithm terminates generating partial combinations when two 1following partial combinations have their corresponding sum of weights below the level of c. At each step of the algorithm, all possible full combinations $(n_1, \ldots n_{k-1}, n_k)$ are generated from the current partial combination $(n_1, \ldots n_{k-1})$. If for the full combination its corresponding sum of weights is greater or equal c $(\sum_{i=1}^{k} n_i \cdot l_i \ge c)$, then we increment the value $t_c(n)$ by $\prod i = 1^k \binom{m_i}{n_i}$, where $n = n_1 + \ldots + n_k$. As the outcome, we obtain $t_c(n)$ and, thus, also $\mathbb{P}(X \ge c|N = n)$ for each n.

The last step is to calculate formula 1.12 using these calculated probabilities.

Note that, since support is a special case of weighted support, the same formula and algorithm can be used to compute its corresponding p-values.

1.2.2.6 Tree-generalized hypergeometric score

As a part of this work, we also introduce a new association score, called tree-generalized hypergeometric (TGH) score, which is conceptually similar to the

CCTSWEEP score proposed by Habib et al. (2007).

We consider a set of bacterial strains $\mathcal{S}$ with its rooted phylogenetic tree T, whose leaves correspond to the strains in $\mathcal{S}$. Let V_T denote the set of all nodes (internal and leaves) in T. Let additionally, function $P_T : V_T \Rightarrow V_T \cup \{\text{null}\}$, for a given $\omega \in V_T$, return its parent node; or null for the root node. Let also function C_T, for a given node $\omega \in V_T$, return the set of its immediate child nodes.

We also introduce function L_T which, for each node ω in T, returns the subtree of descendants of the node, including the node itself. We say these nodes are visible from ω. Additionally, the function L_T applied to any subset c of V_T returns the union of all nodes visible from nodes in the set. More formally, $L_T(c) = \bigcup_{\omega \in V_T} L_T(\omega)$.

In order to present the formal definition of TGH, we first define some auxiliary concepts.

Let $\overline{r} : V_T \rightarrow \{\text{'?'}, \text{'S'}, \text{'R'}\}$ denote the *tree-extended resistance profile* defined recursively as follows:

$$\overline{r}(\omega) = \begin{cases} r(S) & \text{if } \omega \text{ is a leaf node corresponding to strain } S \\ \text{'S'} & \exists_{\omega' \in C_T(\omega)} \overline{r}(\omega') = \text{'S'} \\ \text{'R'} & \neg\exists_{\omega' \in C_T(\omega)} \overline{r}(\omega') = \text{'S'} \wedge \exists_{\omega' \in C_T(\omega)} \overline{r}(\omega') = \text{'R'} \\ \text{'?'} & \text{otherwise} \end{cases} \tag{1.14}$$

Analogously, let $\widehat{b} : V_T \rightarrow \{\text{'?'}, \text{'0'}, \text{'1'}\}$ denote the *tree-extended binary mutation profile* defined recursively as follows:

$$\widehat{b}(\omega) = \begin{cases} b(S) & \text{if } \omega \text{ is a leaf node corresponding to strain } S \\ \text{'0'} & \exists_{\omega' \in C_T(\omega)} \widehat{b}(\omega') = \text{'0'} \\ \text{'1'} & \neg\exists_{\omega' \in C_T(\omega)} \widehat{b}(\omega') = \text{'0'} \wedge \exists_{\omega' \in C_T(\omega)} \widehat{b}(\omega') = \text{'1'} \\ \text{'?'} & \text{otherwise} \end{cases} \tag{1.15}$$

For a given tree T, we call a subset c of its nodes a coloring, if it satisfies the following two conditions:

(A) each path from a leaf to the root contains at most one node from c,

(B) each internal node in T has at least one immediate child node which does not belong to c.

We call a coloring $\overline{c}$ *induced* by a given drug resistance profile r, if it contains the set of nodes in which drug resistance was acquired. More formally, we define a coloring induced by a drug resistance profile r, using its corresponding tree-extended resistance profile $\overline{r}$, as:

$$\overline{c} = \{\omega \in V_T : \overline{r}(\omega) = \text{'R'} \wedge \big(P_T(\omega) = \text{null} \vee \overline{r}(P_T(\omega)) = \text{'S'}\big)\}. \tag{1.16}$$

Analogously, we call a coloring $\widehat{c}$ induced by a given binary mutation profile b, if it contains the set of nodes in which the mutation was acquired. More formally, we define a coloring induced by a binary mutation profile b, using its corresponding tree-extended mutation profile $\widehat{b}$, as:

$$\widehat{c} = \{\omega \in V_T : \widehat{b}(\omega) = \text{'1'} \wedge \Big(P_T(\omega) = \text{null} \vee \widehat{b}(P_T(\omega)) = \text{'0'}\Big)\}. \tag{1.17}$$

Figure 1.3 (A) presents an example of colorings induced by a given drug resistance profile (large red nodes) and a given binary mutation profile (small orange nodes) for a flat tree. Figure 1.3 (B) presents another example of colorings induced by the same pair of profiles, but for a tree which is not flat. In this model the dependencies between different strains are captured by the topology of the tree.

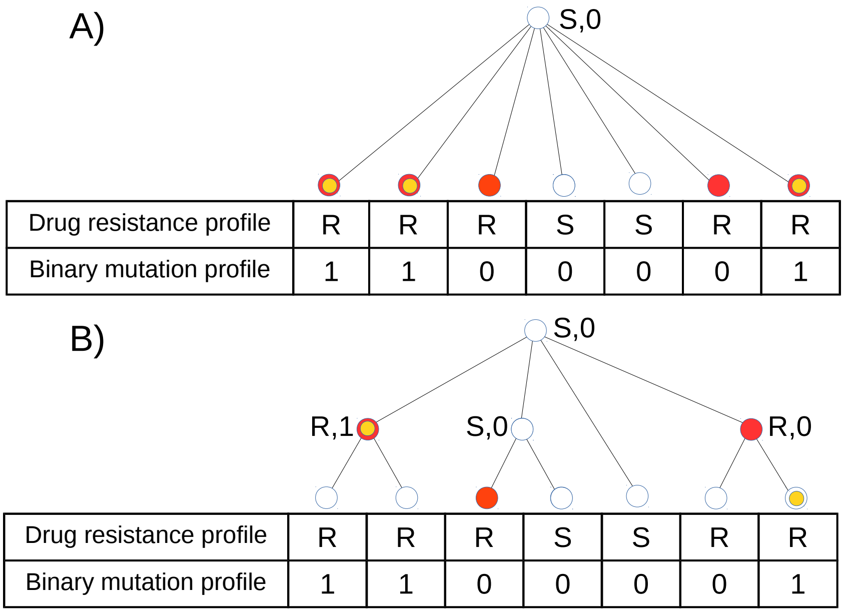
$$W_\omega(n) = \#\{c \in \mathcal{C}_T(\omega) : |c| = n\} \tag{1.18}$$

Here, $\mathcal{C}_T(\omega)$ denotes the set of all colorings of $L_T(\omega)$. We denote by $W_T(n)$, the value of $W_\omega(n)$ for the root node ω in T.

We also define $B_{\omega,\overline{c}}(k,n)$ as the number of colorings of size n, such that exactly k nodes of that coloring are visible from nodes of coloring $\overline{c}$. More formally,

$$B_{\omega,\overline{c}}(k,n) = \#\{c \in \mathcal{C}_T(\omega) : |L_T(\overline{c}) \cap c| = k \wedge |c| = n\} \tag{1.19}$$

We denote by $B_{T,\overline{c}}(k,n)$ the value of $B_{\omega,\overline{c}}(k,n)$ for the root node ω in T.



Drug resistance profile	R	R	R	S	S	R	R
Binary mutation profile	1	1	0	0	0	0	1

Drug resistance profile	R	R	R	S	S	R	R
Binary mutation profile	1	1	0	0	0	0	1

Figure 1.3: **(A)** an example of a pair of a drug resistance profile and a binary mutation profile. Values of the corresponding tree-extended binary mutation profile, and the corresponding tree-extended drug resistance profile are shown next to the nodes. Nodes belonging to the coloring induced by the drug resistance profile $\overline{c}$ are indicated by large red nodes, whereas nodes belonging to the coloring induced by the binary mutation profile $\widehat{c}$ are indicated by small orange nodes. In this example $|\overline{c}| = 5$, $|\widehat{c}| = 3$ and $|L_T(\overline{c}) \cap \widehat{c}| = 3$. **(B)** colorings $\overline{c}$ and $\widehat{c}$ induced by the same pair of profiles but for a different tree. In this example $|\overline{c}| = 3$, $|\widehat{c}| = 2$ and $|L_T(\overline{c}) \cap \widehat{c}| = 2$.

Finally, for a drug resistance profile r and a binary mutation profile b, we denote the colorings induced by the profiles as $\overline{c}$ and $\widehat{c}$, respectively. Let additionally, $n = |\overline{c}|$ and $k = |L(\overline{c}) \cap \widehat{c}|$. Then, we finally define the TGH score, as follows:

$$\mathrm{TGH}_T(r, b) = -log\Big(\frac{\sum_{i=k}^{n} B_{T,\overline{c}}(i, n)}{W_T(n)}\Big). \tag{1.20}$$

We take the negative logarithm to have consistent property, with other scoring methods, such that the higher the score the more likely drug resistance profile r is associated with binary mutation profile b.

The algorithm for TGH Here we describe the algorithm we use to compute the TGH score for a set of pairs of drug resistance profiles and binary mutation profiles.

Naturally, for each leaf node ω in T, two colorings exist: $c_1 = \{\omega\}$, $c_2 = \emptyset$. The following lemma 1 characterizes colorings for internal nodes of T.

Lemma 1. *Let ω be an internal node in T with l immediate child nodes $(\omega_1, \dots \omega_l)$. Let c be a subset of V_T. Then, c is a coloring of $L_T(\omega)$ if and only if $c = \{\omega\}$ or $\big(\omega \notin c$ and $c \neq \{\omega_1, \dots \omega_l\}$ and $c \cap L_T(\omega_i)$ is a coloring of $L_T(\omega_i)$, for each $\omega_i\big)$.*

Proof. $\Rightarrow$: Proof by contradiction. Let us assume $c \neq \{\omega\}$. If $c = \{\omega_1, \dots \omega_l\}$, then c contradicts with the (B) condition of the definition of a coloring. Thus, there exists ω_i, such that, $c \cap L_T(\omega_i)$ does not satisfy (A) or (B). Since $c \cap L_T(\omega_i)$ is a subset of c, c also violates the corresponding (A) or (B) condition. Hence, it contradicts with our assumption that c is a coloring.
$\Leftarrow$: naturally, $\{\omega\}$ satisfies both (A) and (B). Otherwise, ince $\omega \notin c$, $c = \bigcup_{\omega_i} c \cap L_T(\omega_i)$. Thus, c satisfies (A). The condition (B) is satisfied unless $c \cap L_T(\omega_i) = \{\omega_i\}$, but this case is excluded as a separate case. □

Based on the proposition 1 we can derive the following recursive formulas for $W_\omega(n)$. If ω is a leaf node in T, then:

$$W_\omega(n) = [n = 0] + [n = 1] \tag{1.21}$$

If ω is an internal node in T, then:

$$W_\omega(n) = \underbrace{[n = 1]}_{c=\{\omega\}} - \underbrace{[n = l]}_{\{\omega_1,\dots,\omega_l\} \text{ is not a coloring}} + \sum_{\substack{0\le n_1 \le n,\dots,0\le n_l\le n \\ n_1+\dots+n_l=n}} \prod_{i=1}^{l} W_{\omega_i}(n_i) \tag{1.22}$$

Similarly, we can derive the recursive formulas for $B_{\omega,\overline{c}}(k,n)$. If ω is a leaf node in T, then:

$$\begin{aligned} B_{\omega,\overline{c}}(k,n) = \ & [n = 1 \wedge k = 1 \wedge \overline{c} = \{\omega\}] \\ & +[n = 1 \wedge k = 0 \wedge \overline{c} \neq \{\omega\}] \\ & -[n = l \wedge k = |L(\overline{c}) \cap \{\omega_1, \dots, \omega_l\}|] \end{aligned} \tag{1.23}$$

If ω is an internal node in T, then:

$$\begin{aligned} B_{\omega,\overline{c}}(k,n) = \;& [n = 1 \wedge k = 1 \wedge \overline{c} = \{\omega\}] \\ & + [n = 1 \wedge k = 0 \wedge \overline{c} \neq \{\omega\}] \\ & - [n = l \wedge k = |L(\overline{c}) \cap \{\omega_1, \ldots, \omega_l\}|] \\ & + \sum_{\substack{0 \leq n_1 \leq n, \ldots, 0 \leq n_l \leq n \\ n_1 + \ldots + n_l = n \\ 0 \leq k_1 \leq n_1, \ldots, 0 \leq k_l \leq n_l \\ k_1 + \ldots + k_l = k}} \prod_{i=1}^{l} B_{\omega_i,\overline{c}}(k_i, n_i) \end{aligned} \tag{1.24}$$

The pseudocode 1 presents the following steps of the algorithm to compute the TGH score for each pair of drug resistance profile and binary mutation profile. These steps, for a given drug resistance profile r, comprise: (i) simplification of the input tree T' to T by removal of the leaves corresponding to the strains with unknown drug resistance status (according to r); (ii) computation of the tree-extended resistance profile $\overline{r}$ and its corresponding coloring $\overline{c}$; (iii) computation of the values of $W_\omega(n)$ for each n and $\omega \in V_T$, following the recursive formulas 1.21 and 1.22 from the leaves to the root (dynamic programming technique); (iv) computation of the values of $B_{\omega,\overline{c}}(k,n)$ for each k, n and $\omega \in V_T$, following the recursive formulas 1.23 and 1.24, from the leaves to the root (dynamic programming technique); (from leaves to the root); (v) for each binary mutation profile $b \in \mathcal{B}$, computation of the tree-extended binary mutation profile $\widehat{b}$ and its corresponding coloring $\widehat{c}$; and finally (vi) computation of the TGH score based on formula 1.20.

Additionally, in order to speed up the computations of the and $W_T(n)$ and $B_{T,\overline{c}}(k,n)$ values we use the memorization technique to cache results depending on the topology of a subtree. The subtree topologies, used as hashes, are represented as strings in the Nawick tree format enriched by the additional information of belonging to $\overline{c}$, for each node.

Also, due to high time complexity of the score with respect to the maximal number of immediate children of a node, in all computational experiments we calculate the actual TGH score as an average over TGH scores obtained for trees generated by randomly binarizing the input tree.

Algorithm 1 Pseudocode for computing the TGH score

```
Require: A set $\mathcal{S}$ of bacterial strains; with a phylogenetic tree $T'$, a set of binary
    resistance profiles $\mathcal{R}$, and a set of binary mutation profiles $\mathcal{B}$. The function
    $simplify$ removes a node $\omega$ from the tree $T'$ if the strains corresponding to
    the set of leaves visible from $\omega$ have all unknown drug resistance status in $r$.
    After this step, it removes all internal nodes of degree one.
for all $r \in \mathcal{R}$ do
    $T \leftarrow simplify(r, T')$
    compute the tree-extended resistance profile $\overline{r}$ for $r$ in $T$
    compute the coloring $\overline{c}$ induced for $\overline{r}$ in $T$
    compute $W_\omega(n)$ bottom-up for every $n$ and $\omega \in V_T$ , following the 1.21 and
        1.22 formulas
    compute $B_{\omega,\overline{c}}(k, n)$ bottom-up for every $k$, $n$ and $\omega \in V_T$, following the
        1.23 and 1.24 formulas
    for all $b \in \mathcal{B}$ do
        compute the tree-extended mutation profile $\widehat{b}$ for $b$ in $T$
        compute the coloring $\widehat{c}$ for $\widehat{b}$ in $T$
        $n \leftarrow |\widehat{c}|$
        $k \leftarrow |L_T(\overline{c}) \cap \widehat{c}|$
        TGH $\leftarrow -log\left(\frac{\sum_{i=k}^{n} B_{T,\overline{c}}(i,n)}{W_T(n)}\right)$
    end for
end for{These computations are done in parallel for each drug resistance
    profile $r \in \mathcal{R}$}
return TGH score for each pair $r \in \mathcal{R}$ and $b \in \mathcal{B}$.
```

1.2.2.7 Rank-based metascore

Finally, we introduce an association score, called *Rank-based metascore* (RBM), which combines a set of scores into a new score. This approach is based on the natural assumption that each individual score has its own good and weak points. Thus, RBM tries to compromise between the different approaches used to define different scores. This score is based on rankings after sorting with accordance to the scores being combined, rather than the absolute values of the scores.

Let $S_1, S_2, \ldots, S_k$ denote the set of different scores to be combined with RBM. Then, for a given binary mutation profile $b \in \mathcal{B}$ and resistance profile $r \in \mathcal{R}$, the score is defined as the sum of average rankings of b with accordance to scores in question. More formally,

$$\mathrm{RBM}(S_1, \ldots, S_k)(b, r) = \sum_{i=1}^{k} \frac{\mathrm{rank}_u^{S_i}(b, r) + \mathrm{rank}_d^{S_i}(b, r)}{2}. \tag{1.25}$$

Here, $\mathrm{rank}_u^{S_i}(b, r)$ denote the highest ranking of the binary mutation profile with the same S_i score as b, which is the number of binary mutation profiles with the S_i score higher than b plus 1, more formally:

$$\mathrm{rank}_u^{S_i}(b, r) = \#\{b' \in \mathcal{B} : S_i(b', r) > S_i(b, r)\} + 1. \tag{1.26}$$

Analogously, we define $\mathrm{rank}_d^{S_i}(b, r)$ as the lowest ranking of the binary mutation profile with the same S_i score as b, which is the number of binary mutation profiles with the S_i score higher or equal than b, more formally:

$$\mathrm{rank}_d^{S_i}(b, r) = \#\{b' \in \mathcal{B} : S_i(b', r) \geq S_i(b, r)\}. \tag{1.27}$$

Note that, if each binary mutation profile has a different score, the formula $\frac{\mathrm{rank}_u^{S_i}(b,r)+\mathrm{rank}_d^{S_i}(b,r)}{2}$ simplifies to return the ranking of b on the sorted list of binary mutation profiles with respect to the score S_i.

In order to compute the RBM, assuming all the individual scores are already computed, we sort the lists of mutations for each individual score and drug resistance profile r, separately. Then we compute the rank_u and rank_d mappings. Finally, we compute the actual RBM.

Note that, unlike the other scores presented in this work, here, the lower the value of the score the higher the chance the association is real. This definition of RBM is consistent with the current implementation of the score.

In the thesis we consider three versions of the score: (i) combining all the tree-ignorant scores, denoted: RBM (MI,OR,H); (ii) combining WS and TGH, denoted RBM (WS,TGH); and combining (iii) all the five individual scores, denoted RBM (MI,OR,H,WS,TGH) and also shortly RBM (ALL). Note that RBM (MI,OR,H) can be categorized as tree-ignorant score, whereas RBM (WS,TGH) and RBM (ALL) as tree-aware.

1.2.3 Time complexity

Let D denote the number of drug resistance profiles considered. Additionally, let N denote the number of considered strains and M denote the number of binary mutation profiles. Finally, let K denote the maximal number of children of an internal node in the tree. Then, the time complexity of the algorithms we implemented to compute the hypergeometric score, the mutual information, odds ratio, and weighted support is $O(D \cdot N \cdot M)$.

In order to compute the TGH score for the input tree T, based on the formulas 1.23 and 1.24, we implement the dynamic programming algorithm to compute bottom-up the values $B_{\omega,\overline{c}}(k,n)$ for each internal node ω in T, k and n. The time complexity of computing these values for all the nodes is $O(\cdot N^{2\cdot(K-1)} \cdot N)$. Similarly, based on the recursive formulas 1.21 and 1.22,we implement the dynamic programming algorithm to compute bottom-up the values $W_\omega(n)$ for all nodes in T and n. The time complexity of this step is $O(\cdot N^{\cdot(K-1)} \cdot N)$.

This strategy gives the algorithm to compute the TGH score with time complexity $O(D \cdot N^{2\cdot(K-1)} \cdot N + D \cdot N \cdot M)$ which simplifies to $O(D \cdot N \cdot (M + N^2))$ for binary trees.

The time complexity of the algorithm to compute the RBM for a set of E individual scores, assuming the scores are already computed, is $O(D \cdot E \cdot M)$. Note that the time complexity does not depend on the number of strains N.

1.3 Results and Discussion

Here we present the results of applying GWAMAR to three datasets. One for *S. aureus* and two for *M. tuberculosis*.

1.3.1 S. aureus dataset

We first discuss the computational experiment on the dataset of 100 *S. aureus* strains. We use this case study to show the usability of GWAMAR to identify genes associated with drug resistance.

1.3.1.1 Genotype data

We collected genotype data (genome sequences and annotations) for 100 fully sequenced strains of *S. aureus* from the GenBank (Benson et al., 2013) and PATRIC databases (Gillespie et al., 2011). Additionally, genotype data for strain *EMRSA-15* were downloaded from the Wellcome Trust Sanger Institute website. At the time of writing, 31 out of the 100 *S. aureus* strains had the sequencing status “completed”. For the remaining strains whose genomes were still being assembled, contig sequences (covering around 90% of the genomes) and annotations were used.

We unified the original genome annotations employing CAMBer. However, in order to determine gene families we additionally extended the multigene consolidation graph by edges coming from BLAST amino-acid queries. More formally, we added an edge between a pair of genes to the consolidation graph if the percent of identity (calculated as the number of identities over the length of the longer gene) of the BLAST hit between them exceeded a threshold $P(L)$ given by the HSSP curve formula (Rost, 1999):

$$P(L) = \begin{cases} 100 & L \leq 11 \\ c + 480 \cdot L^{-0.32 \cdot (1+e^{-L/1000})} & 11 < L \leq 450 \\ c + 19.5 & L > 450 \end{cases} \tag{1.28}$$

Here, c was set to 40.5 and L is the number of aligned amino-acid residues.

Then, each connected component in the multigene consolidation graph corresponds to a gene family. We computed multiple alignments using MUSCLE (Edgar, 2004) for all these gene families.

In this work, unlike in the current version of GWAMAR, we considered two kinds of genetic variations (mutations):

- gene gain/losses,
- point mutations (in amino-acid sequences).

In comparison to the current version of GWAMAR, we did not take into account mutations in gene promoter regions. Here, point-mutation profiles are transformed from columns in multiple alignments computed for gene families with elements present in at least $|\mathcal{S}| - 1$ strains.

1.3.1.2 Phylogenetic tree of the strains

We computed the phylogenetic tree of the input strains using a consensus method with majority rule implemented in the PHYLIP package, developed by Felsenstein (2005). We applied the consensus method to trees constructed for all gene families with exactly one element in each strain. The trees were constructed using the maximum likelihood approach implemented in the PHYLIP package.

1.3.1.3 Phenotype data (drug susceptibility)

We performed a careful search of the literature for results of drug susceptibility tests of the strains considered. The drug susceptibility data were collected from the following sources: (i) 25 publications issued together with the fully sequenced genomes; (ii) NARSA project (`http://www.narsa.net`); (iii) email exchange with the authors of publications related to strains *ST398* and *TW20*; and (iv) other publications found by searching related literature. In total we used 71 publications to retrieve the drug resistance information.

1.3.1.4 Assessment of accuracy

We verified the usability of our approach by trying to re-identify known drug resistance determinants. In this experiment, we compared the proposed scoring — support and weighted support— to odds ratio, which is a popular measure used in genome-wide association studies. Table 1.1 shows rankings of the gene-gain/-loss profiles for genes which are known drug resistance determinants. The experiment suggests that weighted support outperforms both: support and odds ratio. The latter two scores do not incorporate additional information about phylogeny

1.3.1.5 Prediction of resistance

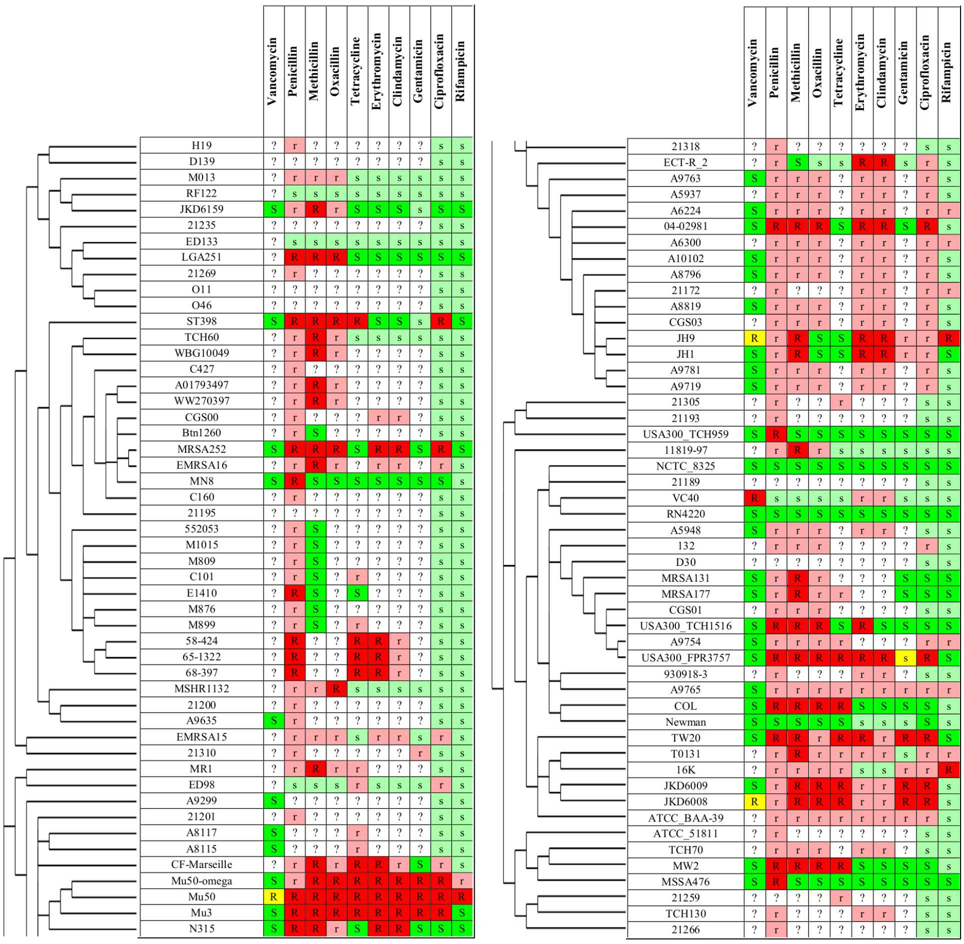
This experiment also reveals that the amount of the collected drug resistance information is not sufficient to correctly identify drug resistance-associated genes. However, the high consistency of drug resistance profiles corresponding to the collected information and the presence of drug resistance determinants (summing over drugs, there are 117 drug-resistant strains, where only 4 of them do not

		Rankings before prediction			**Rankings after prediction**		
gene id.	drug name	S	WS	OR	S	WS	OR
tet	tetracycline	54.5	2.5	43.7	1.5	1.5	1.5
tetM	tetracycline	14.5	11.5	7.5	4	4	4
mecA	methicillin	1	1	1	1	1	1
mecA	oxacillin	3	4	2	1	2	1
ermA1	clindamycin	5.5	5.5	5.5	1	1	1
ermC	clindamycin	907	471	907	414.5	11	191.5
ermA1	erythromycin	3	3	4	1	1	1
ermC	erythromycin	1527	3994.5	1006.5	413.5	28	214.5
aacA-aphD	gentamicin	72	34	34	1	1	1
blaZ	penicillin	163	66	223	1.5	1	2.5
mecA	penicillin	163	8	223	11	5	52
Average ranking (excluding ermC):		53.27	**15.05**	60.411	2.55	**1.94**	7.22

Table 1.1: Rankings of the known drug resistance determinants obtained by employing three different methods to score gene-gain/-loss profiles: support (S), weighted support (WS) and odds ratio (OR). Since some of the gene-gain/-loss profiles are assigned with the same score, we calculate their rankings as the arithmetic mean of positions of the profiles with the same score on the list sorted according to the scores; thus some of the rankings are not round numbers. The rankings were computed before and after prediction of drug resistance, which is based on the presence of drug resistance determinants. We excluded the gene *ermC* from the calculations of average rankings since none of the methods were able to pull it out into the top 100 before prediction.

have any known drug resistance determinants; and there are 112 drug-susceptible strains, where only 8 of them have at least one drug resistance determinant) suggests that we can use the determinants to predict drug resistance in the strains without drug resistance information available.

It is perhaps questionable to predict drug resistance in those strains for which the whole-genome sequence is not determined yet. So we did prediction only for those strains with completed sequencing or at least information on their plasmids (which often carry the drug resistance determinants). Nevertheless, we predicted drug resistance also for those strains that were not yet fully sequenced, provided the presence of drug resistance-determining genes had been confirmed for them. Moreover, we predicted drug resistance to rifampicin and ciprofloxacin for all 100 strains, as the drug resistance for rifampicin and ciprofloxacin is determined by point mutations in genes *rpoB*, *gyrA* and *grlA* (synonymous name to *parC*), which were sequenced in all strains. More precisely, we predicted as rifampicin-resistant all strains with any mutation present in the rifampicin resistance determining region (RRDR). We defined the RRDR as the amino-acid range from 463 to 530 in the *rpoB* gene sequence (according to (O'Neill et al., 2006)). Analogously, we predicted as ciprofloxacin-resistant all strains with any point mutation in the quinolone resistance determining region (QRDR). We defined QRDR as the amino-acid ranges from position 68 to 107 and from position 64 to 103 in the *grlA* and *parC* gene sequences, respectively (according to (Ferrero et al., 1995)). Figure 1.4 shows the complete information about drug susceptibility after prediction.



	Vancomycin	Penicillin	Methicillin	Oxacillin	Tetracycline	Erythromycin	Clindamycin	Gentamicin	Ciprofloxacin	Rifampicin
H19	?	r	?	?	?	?	?	?	s	s
D139	?	?	?	?	?	?	?	?	s	s
M013	?	r	r	r	s	s	s	s	s	s
RF122	?	s	s	s	s	s	s	s	s	s
JKD6159	S	r	R	r	S	S	S	s	S	S
21235	?	?	?	?	?	?	?	?	s	s
ED133	?	s	s	s	s	s	s	s	s	s
LGA251	?	R	R	R	S	S	S	S	S	S
21269	?	r	?	?	?	?	?	?	s	s
O11	?	?	?	?	?	?	?	?	s	s
O46	?	?	?	?	?	?	?	?	s	s
ST398	S	R	R	R	R	S	S	s	R	S
TCH60	?	r	R	r	s	s	s	s	s	s
WBG10049	?	r	R	r	?	?	?	?	s	s
C427	?	r	?	?	?	?	?	?	s	s
A01793497	?	r	R	r	?	?	?	?	s	s
WW270397	?	r	R	r	?	?	?	?	s	s
CGS00	?	r	?	?	?	r	r	?	s	s
Btn1260	?	r	S	?	?	?	?	?	s	s
MRSA252	S	R	R	R	S	R	R	S	R	S
EMRSA16	?	r	R	r	?	r	r	?	r	s
MN8	S	R	S	S	S	S	S	S	S	s
C160	?	r	?	?	?	?	?	?	s	s
21195	?	?	?	?	?	?	?	?	s	s
552053	?	r	S	?	?	?	?	?	s	s
M1015	?	r	S	?	?	?	?	?	s	s
M809	?	r	S	?	?	?	?	?	s	s
C101	?	r	S	?	r	?	?	?	s	s
E1410	?	R	S	?	S	?	?	?	s	s
M876	?	r	S	?	?	?	?	?	s	s
M899	?	r	S	?	r	?	?	?	s	s
58-424	?	R	?	?	R	R	r	?	s	s
65-1322	?	R	?	?	R	R	r	?	s	s
68-397	?	R	?	?	R	R	r	?	s	s
MSHR1132	?	r	r	R	s	s	s	s	s	s
21200	?	r	?	?	?	?	?	?	s	s
A9635	S	r	?	?	?	?	?	?	s	s
EMRSA15	?	r	r	r	s	r	r	s	r	s
21310	?	r	?	?	?	?	?	r	s	s
MR1	?	r	R	r	r	?	?	?	s	s
ED98	?	s	s	s	r	s	s	s	r	s
A9299	S	?	?	?	?	?	?	?	s	s
21201	?	r	?	?	?	?	?	?	s	s
A8117	S	?	?	?	r	?	?	?	s	s
A8115	S	?	?	?	r	?	?	?	s	s
CF-Marseille	?	r	R	r	R	R	r	S	r	s
Mu50-omega	S	r	R	R	R	R	R	R	R	r
Mu50	R	R	R	R	R	R	R	R	R	R
Mu3	S	R	R	R	R	R	R	R	R	S
N315	S	R	R	r	S	R	R	S	S	S
21318	?	r	?	?	?	?	?	?	s	s
ECT-R_2	?	r	S	s	s	R	R	s	r	s
A9763	S	r	r	r	?	r	r	?	r	s
A5937	?	r	r	r	?	r	r	?	r	s
A6224	S	r	r	r	?	r	r	?	r	r
04-02981	S	R	R	R	S	R	R	S	R	s
A6300	?	r	r	r	?	r	r	?	r	r
A10102	S	r	r	r	?	r	r	?	r	s
A8796	S	r	r	r	?	r	r	?	r	s
21172	?	r	?	?	?	r	r	?	r	r
A8819	S	r	r	r	?	r	r	?	r	s
CGS03	?	r	r	r	?	r	r	?	r	s
JH9	R	r	R	S	S	R	R	r	r	R
JH1	S	r	R	S	S	R	R	r	r	S
A9781	S	r	r	r	?	r	r	?	r	s
A9719	S	r	r	r	?	r	r	?	r	s
21305	?	r	?	?	r	?	?	?	s	s
21193	?	r	?	?	?	?	?	?	s	s
USA300_TCH959	S	R	S	S	S	S	S	S	S	S
11819-97	?	r	R	r	s	s	s	s	s	s
NCTC_8325	S	S	S	S	S	S	S	S	S	S
21189	?	?	?	?	?	?	?	?	s	s
VC40	R	s	s	s	s	r	r	s	s	s
RN4220	S	S	S	S	S	S	S	S	S	S
A5948	S	r	r	r	?	r	r	?	s	s
132	?	r	r	r	?	?	?	?	r	s
D30	?	?	?	?	?	?	?	?	s	s
MRSA131	S	r	R	r	?	?	?	S	S	S
MRSA177	S	r	R	r	r	?	?	S	S	S
CGS01	?	r	r	r	?	?	?	?	s	s
USA300_TCH1516	S	R	R	R	S	R	S	S	S	S
A9754	S	r	r	r	r	?	?	?	r	r
USA300_FPR3757	S	R	R	R	R	R	R	s	R	S
930918-3	?	r	?	?	?	r	r	?	s	s
A9765	S	r	r	r	r	r	r	r	r	r
COL	S	R	R	R	R	S	S	S	S	s
Newman	S	S	S	S	S	s	s	s	S	s
TW20	S	R	R	r	R	R	r	R	R	S
T0131	?	r	R	r	r	r	r	s	r	r
16K	?	r	r	r	r	s	s	r	r	R
JKD6009	S	r	R	R	R	r	r	R	R	s
JKD6008	R	r	R	R	R	r	r	R	R	s
ATCC_BAA-39	?	r	r	r	r	r	r	r	r	s
ATCC_51811	?	r	?	?	?	?	?	?	s	s
TCH70	?	r	r	r	?	r	r	?	s	s
MW2	S	R	R	R	R	S	S	S	S	s
MSSA476	S	R	S	S	S	S	S	S	S	S
21259	?	?	?	?	r	?	?	?	s	s
TCH130	?	r	?	?	?	r	r	?	s	s
21266	?	r	?	?	?	?	?	?	s	s

Figure 1.4: The collected dataset of phenotypes put together with results of our drug resistance predictions based on the presence of known drug resistance determinants. Due to the high number of strains the table is split into two panels. Columns represent drugs, rows represent *S. aureus* strains included in the study in the order corresponding to the reconstructed phylogenetic tree of strains. Green, yellow and red cell colors represent susceptible, intermediate resistant and resistant phenotypes, respectively. Analogously, light green and light red cell colors represent predicted susceptible and resistant phenotypes, respectively. White cell color represents unknown (not determined by experiments or prediction) drug resistance phenotypes.

1.3.1.6 Essential mutations

Here, we distinguish two categories of gene-gain/-loss and point-mutation profiles depending on how they correspond to a given drug resistance profile. We categorize a given mutation profile m as:

- *Essential* mutation, when m is absent in all drug-susceptible strains,
- *Conflict* mutation, when m is present in at least one drug-susceptible strain.

Further, we distinguish *neutral* mutations as a subclass of essential mutations, these are essential mutations that are not present in any of drug-resistant strains. Thus, neutral mutations may only be present in strains with unknown drug-resistance status.

Analogously, we transfer the above introduced concepts to gene-gain/-loss profiles, defining essential, neutral and conflict gene-gain/-loss profiles.

1.3.1.7 Detection of drug resistance-associated mutations

Then, we applied our approach to the dataset supplemented by the predicted information about drug susceptibility for the following drugs: tetracycline, β-lactams (penicillin, oxacillin, methicillin), erythromycin, gentamicin, vancomycin, ciprofloxacin and rifampicin.

Below we discuss the results of our approach applied separately to the following drugs: tetracycline, β-lactams (penicillin, methicillin), erythromycin, gentamicin, vancomycin, ciprofloxacin. We do not discuss here results for oxacillin and clindamycin, since they have very similar drug resistance profiles to methicillin and erythromycin, respectively. All other drugs were excluded from the analysis due to low number of strains with available drug resistance information on these drugs.

Tables 1.2 and 1.3 present the top-scored gene-gain/-loss, and point-mutation profiles for the discussed drugs, respectively. The genes presented in the tables were selected according to the following procedure: for each drug we construct a function, which gives for each gene (listed in descending order with respect to normalized weighted support) minus logarithm of p-value ($-log$(p-value)) of this score. Then, we report genes which correspond to the portion of the graph of this function before it gets flattened.

Gene identifier	NS	NWS	OR	p-value	Gene functional annotation
Penicillin (NWS-threshold: 0.58)					
⋆ SAR1831(blaZ)	0.84	0.81	37.15	1.15e-06	beta-lactamase
SAR1829(blaI)	0.84	0.74	37.15	5.24e-06	transcriptional repressor
SAR1830(blaR1)	0.82	0.73	31.27	7.09e-06	beta-lactamase regulatory protein blar1
SAR0056	0.63	0.71	12.13	1.03e-05	conserved hypothetical protein
⋆ SAR0039(mecA)	0.61	0.70	10.94	1.28e-05	methicillin resistance determinant mecA
SAR0060(ccrA)	0.61	0.63	10.94	4.40e-05	resolvase, n-terminal domain protein
SAR0061(yycG)	0.61	0.63	10.94	4.40e-05	putative membrane protein
NWMN 0025	0.57	0.63	9.40	4.41e-05	conserved domain protein
SAR0037(ugpQ)	0.60	0.63	10.39	5.08e-05	glycerophosphoryldiester phosphodiesterase
SAR0038(maoC)	0.60	0.63	10.39	5.08e-05	dehydratase
SAR0057	0.57	0.59	9.40	9.78e-05	conserved hypothetical protein
Methicillin (NWS-threshold: 0.68)					
⋆ SAR0039(mecA)	1.00	1.00	950.00	4.48e-20	methicillin resistance determinant mecA
SAR0037(ugpQ)	0.98	0.94	931.00	6.77e-15	glycerophosphoryldiester phosphodiesterase
SAR0038(maoC)	0.98	0.94	931.00	6.77e-15	dehydratase
SAR0056	0.95	0.85	900.00	7.55e-12	conserved hypothetical protein
SAR0036	0.64	0.80	33.78	5.77e-11	putative membrane protein
SAR0057	0.85	0.75	162.00	6.47e-10	conserved hypothetical protein
SAR0060(ccrA)	0.91	0.73	432.00	1.40e-09	resolvase, n-terminal domain protein
SAR0061(yycG)	0.91	0.73	432.00	1.40e-09	putative membrane protein
MW0028(ebpS)	0.54	0.71	22.30	2.76e-09	hmg-coa synthase
Tetracycline (NWS-threshold: 0.32)					
SAAV b3(repC)	0.54	0.64	27.69	5.70e-08	plasmid replication protein
⋆ SATW20 00660(tet)	0.54	0.64	27.69	5.70e-08	tetracycline resistance protein
SATW20 00670(pre)	0.50	0.50	24.00	3.51e-06	plasmid recombination enzyme type 3
⋆ SATW20 04620(tetM)	0.46	0.37	20.80	7.54e-05	tetracycline resistance protein tetM
SATW20 08990(virE)	0.42	0.37	19.93	7.67e-05	pathogenicity island protein
SATW20 09000	0.42	0.37	19.93	7.67e-05	pathogenicity island protein
SATW20 09010(lipA)	0.42	0.37	19.93	7.67e-05	superantigen-encoding pathogenicity islands
SATW20 04610(thiI)	0.43	0.35	18.00	1.32e-04	putative transcriptional regulator
MW0745(int)	0.25	0.32	8.00	2.28e-04	site-specific recombinase, phage integrase
MW0747	0.25	0.32	8.00	2.28e-04	DNA-binding helix-turn-helix protein
Erythromycin (NWS-threshold: 0.27)					
⋆ SAR0050(ermA1)	0.80	0.58	76.00	1.36e-06	rRNA adenine n-6-methyltransferase
CGSSa03 12660	0.47	0.44	17.19	2.98e-05	conserved hypothetical protein
SAR0054(tnpA1)	0.75	0.39	72.00	8.12e-05	transposase for transposon
SAR1734	0.75	0.39	72.00	8.12e-05	methylase
SAR1736(spc2)	0.75	0.39	72.00	8.12e-05	spectinomycin 9-o-adenylyltransferase
SaurJH9 1711(radC)	0.72	0.38	62.00	8.83e-05	predicted protein
SAUSA300 pUSA030006	0.20	0.35	4.75	1.65e-04	replication and maintenance protein
SAR1737(tnpC2)	0.72	0.34	62.00	1.89e-04	Unknown
SAR1529	0.33	0.33	9.15	2.43e-04	conserved hypothetical protein
SATW20 04860(recF 1)	0.23	0.30	5.52	3.67e-04	recombinational DNA repair ATPase
SAR1738(tnpB2)	0.70	0.29	54.00	4.39e-04	transposase B from transposon Tn554
SauraJ 010100009720	0.23	0.27	5.52	6.60e-04	conserved domain protein
Gentamicin (NWS-threshold: 0.83)					
⋆ SaurJH1 2806(aacA-aphD)	0.83	0.90	150.00	9.38e-11	bifunc. acetyltransferase/phosphotransferase
SaurJH1 2805	0.75	0.83	90.00	2.95e-09	GNAT family acetyltransferase
Ciprofloxacin (NWS-threshold: 0.4)					
SATW20 04610(thiI)	0.35	0.45	36.00	1.33e-07	putative transcriptional regulator
SATW20 04650(cap8J)	0.32	0.40	31.57	8.25e-07	lipoprotein
SATW20 04670(capL)	0.32	0.40	31.57	8.25e-07	putative ATP/GTP-binding protein
SATW20 04780	0.32	0.40	31.57	8.25e-07	conjugation related protein
SATW20 04800	0.32	0.40	31.57	8.25e-07	replication initiation factor
SATW20 04810	0.32	0.40	31.57	8.25e-07	DNA segregation ATPase FtsK/SpoIIIE
SATW20 04830	0.32	0.40	31.57	8.25e-07	conjugative transposon protein

Table 1.2: Summarizing information for the top scored gene-gain/-loss profiles. The consequent columns refer to: gene identifier of the corresponding gene family; normalized support (NS); normalized weighted support (NWS); odds ratio (OR); p-value and the gene functional annotation. Thresholds for weighted support are provided in brackets for each drug.

Tetracycline Tetracycline acts by binding to the 30S ribosomal subunit (16S rRNA and the protein encoded by the gene *rpsS* are its direct targets), preventing

Gene identifier	desc.	NS	NWS	OR	p-value	Gene functional annotation
Penicillin (NWS-threshold: 0.4)						
SAR0023(sasH)	G_{723}D	0.55	0.63	8.51	1.87e-05	virulence-associated cell-wall-anchored protein
SAR0023(sasH)	T_{725}A	0.54	0.62	8.11	2.23e-05	virulence-associated cell-wall-anchored protein
SAR0304	V_{295}I	0.39	0.49	4.48	3.25e-04	acid phosphatase
SAR2791	V_{182}M	0.46	0.46	6.05	5.41e-04	transcriptional regulator, Xre family
SAR2700	N_{493}KD	0.52	0.45	7.72	6.16e-04	ABC transporter permease protein
SAR0233(hmp)	Q_{333}K	0.44	0.44	5.48	7.21e-04	flavohemoprotein
SAR0318(sbnA)	N_{25}HK	0.44	0.43	5.48	8.36e-04	alpha/beta family hydrolase
SAR2664	V_{282}AT	0.44	0.43	5.48	8.36e-04	probable monooxygenase
SAR2779	S_{48}G	0.44	0.43	5.48	8.36e-04	n-hydroxyarylamine o-acetyltransferase
SAR0318(sbnA)	T_{138}IM	0.43	0.43	5.21	8.36e-04	alpha/beta family hydrolase
SAR0318(sbnA)	T_{139}AQ	0.43	0.43	5.21	8.36e-04	alpha/beta family hydrolase
SAR0023(sasH)	A_{749}TG	0.41	0.43	4.96	8.44e-04	virulence-associated cell-wall-anchored protein
SAR0318(sbnA)	R_{130}CG	0.41	0.43	4.96	8.72e-04	alpha/beta family hydrolase
SAR0322(folC)	H_{201}YQE	0.41	0.43	4.96	8.72e-04	possibly adp-ribose binding module
SAR0233(hmp)	K_{323}ET	0.40	0.42	4.71	9.08e-04	flavohemoprotein
SAR2750(icaC)	I_{21}V	0.40	0.42	4.71	9.46e-04	polysaccharide intercellular adhesin biosynthesis
SAR0233(hmp)	S_{309}RN	0.39	0.42	4.48	9.46e-04	flavohemoprotein
Methicillin (NWS-threshold: 0.25)						
SAR0198(oppF)	T_{287}IK	0.10	0.29	2.11	1.41e-04	putative glutathione transporter, ATP-binding
SAR0420	I_{72}F	0.10	0.29	2.11	1.41e-04	membrane protein
SAR2508(sbi)	S_{219}AT	0.10	0.29	2.11	1.41e-04	IgG-binding protein Sbi
SAR2508(sbi)	N_{222}QK	0.10	0.29	2.11	1.41e-04	IgG-binding protein Sbi
SAR2508(sbi)	K_{224}SDN	0.10	0.29	2.11	1.41e-04	IgG-binding protein Sbi
Tetracycline (NWS-threshold: 0.2)						
SAR1840	D_{291}YS	0.18	0.23	5.22	7.09e-04	NAD(FAD)-utilizing dehydrogenases
SAR2336(rpsJ)	K_{57}M	0.29	0.23	9.60	7.32e-04	SSU ribosomal protein S10P (S20E)
SAR0550(rpsL)	K_{113}R	0.36	0.20	13.33	1.14e-03	SSU ribosomal protein S12P (S23E)
Erythromycin (NWS-threshold: 0.2)						
SAR0576	A_{68}EV	0.07	0.21	1.54	8.89e-04	phosphoglycolate phosphatase
Gentamicin (NWS-threshold: 0.21)						
SAR1840	L_{289}IW	0.33	0.29	15.00	1.43e-03	NAD(FAD)-utilizing dehydrogenases
SAR1840	D_{291}YS	0.33	0.29	15.00	1.43e-03	NAD(FAD)-utilizing dehydrogenases
SAR1840	H_{327}RF	0.33	0.29	15.00	1.43e-03	NAD(FAD)-utilizing dehydrogenases
SAR1167(ylmH)	K_{215}N	0.25	0.29	10.00	1.43e-03	RNA-binding S4 domain-containing protein
SAR1167(ylmH)	R_{216}V	0.25	0.29	10.00	1.43e-03	RNA-binding S4 domain-containing protein
SAR1167(ylmH)	V_{217}L	0.25	0.29	10.00	1.43e-03	RNA-binding S4 domain-containing protein
SAR0547(rpoB)	D_{471}YG	0.17	0.21	6.00	4.61e-03	DNA-directed RNA polymerase beta subunit
SAR1833(trmB)	T_{54}IK	0.17	0.21	6.00	4.61e-03	tRNA (guanine46-n7-)-methyltransferase
Ciprofloxacin (NWS-threshold: 0.12)						
SAR1367(grlA)	S_{80}YF	1.00	1.00	2244.00	6.03e-30	topoisomerase IV subunit a
SAR0006(gyrA)	S_{90}AL	0.94	0.88	1056.00	1.92e-18	DNA gyrase subunit a
SAR2449(lytT)	V_{45}I	0.21	0.20	17.11	2.06e-04	transcriptional regulator
SAR1840	L_{289}IW	0.12	0.20	8.80	4.56e-04	NAD(FAD)-utilizing dehydrogenases
SAR1793(thiI)	A_{92}ET	0.09	0.20	6.39	2.06e-04	thiamine biosynthesis protein thiI
SAR2212(murA2)	A_{102}T	0.06	0.20	4.12	2.06e-04	UDP-n-acetylglucosamine 1-carboxyvinyltransferase
SAR1367(grlA)	E_{84}KG	0.26	0.15	23.76	9.40e-04	topoisomerase IV subunit a
SAR0235(pstG__1)	F_{401}LV	0.09	0.13	6.39	2.21e-03	PTS system, maltose and glucose-specific
SAR0400(nfrA)	R_{194}H	0.09	0.13	6.39	2.21e-03	nitroreductase family protein

Table 1.3: Summarizing information for the top scored point mutation profiles, only for essential mutations. The conflict mutations were removed from the table for: tetracycline, erythromycin and gentamicin (for the rest of drugs there were no conflict mutations above the set thresholds). The consequent columns refer to: gene identifier of the corresponding gene family; corresponding position in the multiple alignment and changed amino acids; normalized support (NS); normalized weighted support (NWS); odds ratio (OR); p-value (computed as described in section 1.2.2.5) and the gene functional annotation. Thresholds for weighted support are provided in brackets for each drug.

binding of tRNA to the mRNA-ribosome complex, and thus inhibiting protein synthesis (Knox et al., 2011).

The most common drug resistance mechanism to tetracycline in *S. aureus* is mediated by ribosome protection proteins (RPPs) such as *tet* and *tetM*, which bind to the ribosome complex, thus preventing the binding of tetracycline (Chopra and Roberts, 2001; Connell et al., 2003).

Genes *tet* and *tetM*, mediating this mechanism, cover all tetracycline-resistant strains, except *MW2*. The drug resistance status pf *MW2* may be caused by errors in the drug susceptibility test, errors in sequencing, or by some not-yet-known drug resistance mechanism. The inconsistent information about strain *MW2*'s tetracycline susceptibility and the lack of identified drug resistance determinants suggest that the strain is possibly drug susceptible. In our experiment we initially assumed that the tetracycline resistance information is not available for strain *MW2*.

Our method shows that, besides *tet* and *tetM*, there are a few more genes that have highly scored gene-gain/-loss profiles. Especially interesting are the following genes which are not gained by any of the drug susceptible strains: *repC*, *pre*, *thiI*, *int*, *clfB* (see Table 1.2). There are studies reporting the significance of these *clfB* and *repC* genes in drug resistance (McAleese and Foster, 2003; Werckenthin et al., 1996). Interestingly, the gene *repC* seems to co-evolve with *tet* (highly correlated gene-gain/-loss profiles).

Applying our method to point mutations, we have identified two highly scored (and essential) point mutations in ribosomal complex proteins: $K101R$ in *rpsL* and $K57M$ in *rpsJ*. According to our knowledge, this is the first report on the significance of the point mutations for drug resistance in *S. aureus*. However, mutations in *rpsJ* have been associated with tetracycline resistance in another bacteria *Neisseria gonorrhoeae* (Hu et al., 2005).

Beta-lactams Beta-lactams are a broad class of antibiotics, which possess (by definition) the β-lactam ring in their structure. The ring is capable of binding transpeptidase proteins (also known as Penicillin Binding Proteins — PBPs) (Knox et al., 2011), which are important for synthesis of the peptidoglycan layer of bacterial cell wall. PBPs with attached drug molecules are no longer able to synthesize peptidoglycan, leading to bacterial death (Sabath, 1982). In our case study, we consider three β-lactam antibiotics: penicillin, oxacillin and methicillin. However, since the drug resistance profile and drug resistance mechanisms for oxacillin and methicillin are very similar we discuss results only for methicillin.

There are two common resistance mechanisms to β-lactams in *S. aureus* (Sabath, 1982; Drawz and Bonomo, 2010). The first one is mediated by β-lactamase enzymes, which bind drug molecules and break the β-lactam ring, thus deactivating the drug molecules. This mechanism is effective against penicillin (which is β-lactamase sensitive) and not effective against methicillin and oxacillin (which are β-lactamase resistant) (WHO, 2010). The second β-lactam resistance mechanism is mediated by proteins which are capable of functionally substituting for PBPs, but have much smaller affinity to β-lactam molecules. This mechanism is effective against penicillin, methicillin and oxacillin.

Penicillin In our dataset, all strains resistant to penicillin possess proteins responsible for one of the two mechanisms. More precisely, there are 69 drug-resistant strains (with available drug resistance information), which possess *blaZ*—the standard β-lactamase protein (note that its regulators *blaR1* and *blaI* do not always co-occur). All the remaining penicillin-resistant strains have *mecA*, which is an altered PBP. Table 1.2 provides information about the top-scoring gene-gain/-loss profiles.

Applying our method we, have also identified the uncategorized putative protein, *SAR0056*, as putatively associated with penicillin resistance (see Table 1.2).

Methicillin Applying our approach to gene-gain/-loss profiles we identified (beside *mecA*) genes *ugpQ* and *maoC*. The correlation of gene profiles to the profile of *mecA* and their close proximity on the genomes suggests that these genes co-evolve (see Figure 1.5 for more details). This co-evolution may reflect some important role played by these genes in methicillin resistance. This calls for further study of the role of these two genes in methicillin resistance.

We have also identified a few point mutations that are putatively associated with methicillin resistance. Interestingly, two of the mutations in the top 10 essential mutations according to weighted support ($I72F$ in *SAR0420* and $E208Q/K/D$ in *SAR0436*) are present in cell membrane proteins. This suggests some compensatory mechanism to the presence of *mecA*.

Ciprofloxacin Ciprofloxacin belongs to a broad class of antibiotics, called fluoroquinolones, which are functional against bacteria by binding DNA gyrase subunit A (encoded by *gyrA*) and DNA topoisomerase 4 subunit A (encoded by

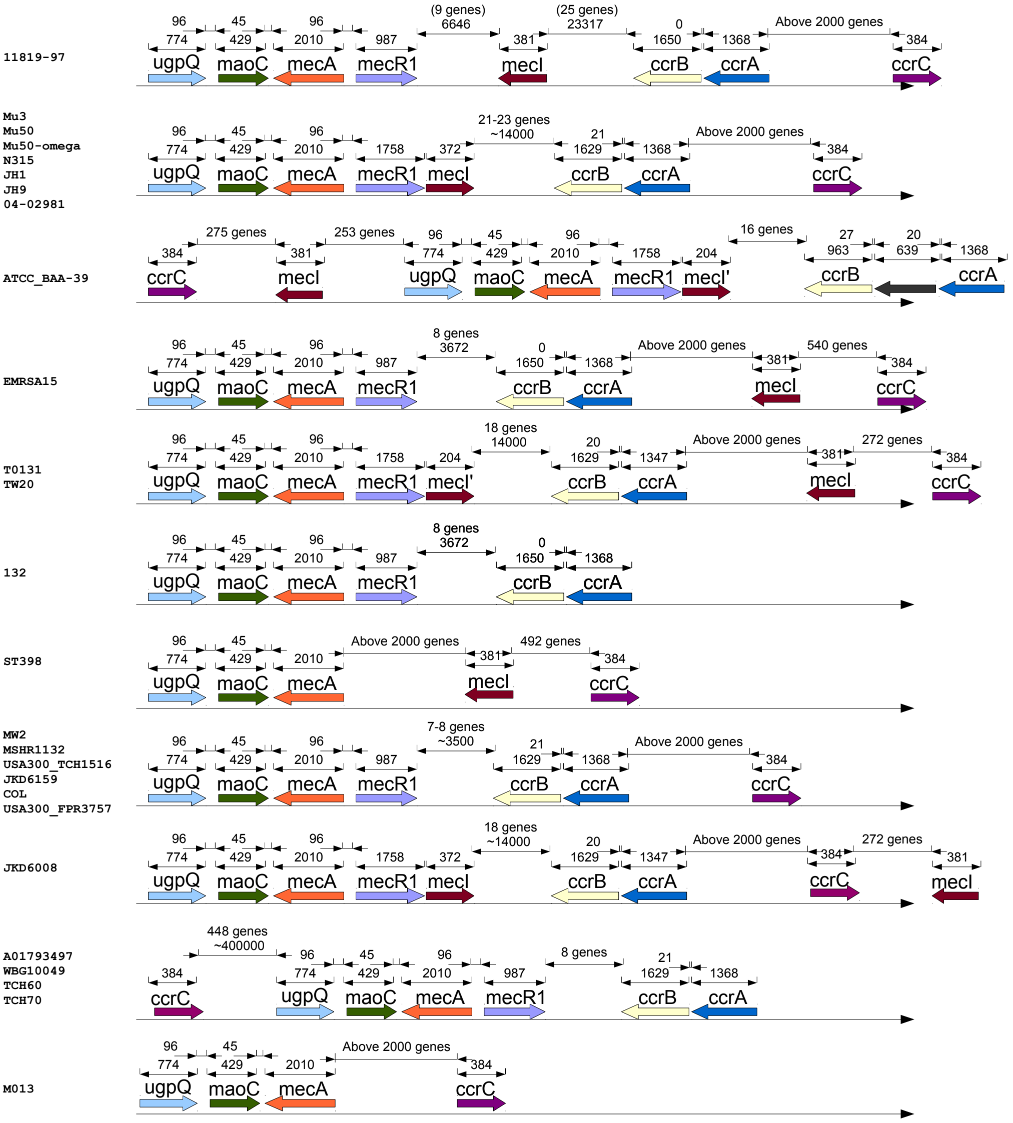


Figure 1.5: Presence and relative genome coordinates of genes related to methicillin resistance (*mecA*, *mecR1*, *mecI*, *ccrA*, *ccrB*, *ccrC*), put together with the identified genes: *ugpQ* and *maoC*. The gene presence profiles are clustered with respect to the genes order. In this figure we include only these methicillin-resistant strains for which all the genes where located on the main genome and within the same sequence contig (in order to determine the relative positions).

parC), which are enzymes necessary to separate bacterial DNA, thereby inhibiting cell division (Knox et al., 2011). The most common ciprofloxacin-resistance mechanism is mediated by point mutations in the drug targets, *parC* and *gyrA*.

Applying our approach we identified (by highest weighted support) two point mutations in ciprofloxacin target genes — $S80F/Y$ in *parC* and $S90A/L$ in *gyrA*— which are located in QRDR and known to be responsible for the first mechanism of ciprofloxacin resistance (Ferrero et al., 1995). The presence of these mutations is correlated with the ciprofloxacin resistance profile for strains with available drug resistance information. However, they differ for two strains *ED98* and *16K* (only the mutation in *parC* is present). This may suggest intermediate drug resistance level for these strains. Unfortunately ciprofloxacin resistance information is not available for these strains.

Erythromycin Erythromycin acts by binding the 23S rRNA molecule (in the 50S subunit) of the bacterial ribosome complex, leading to inhibition of protein synthesis (Knox et al., 2011).

There are three known erythromycin resistance mechanisms (Schmitz et al., 2000). First — the most common mechanism — is by methylation (addition of two residues to the domain V of 23S rRNA) of the 23S rRNA molecule, which prevents the ribosome from binding with erythromycin. This methylation is mediated by enzymes from the *erm* gene family, the most common are *ermA* and *ermC*. The second mechanism is mediated by the presence of macrolide efflux pumps (encoded by *msrA* and *msrB*). The third mechanism is the inactivation of drug molecules by specialized enzymes such as *ereA* or *msrB* (Schmitz et al., 2000).

We found that none of the strains in our case study possess genes *ereA* or *ereB*. Genes encoding efflux pumps (*msrA* and *msrB*) are present also in drug-susceptible strains (for example, *NCTC 8325* and *Newman*), which may suggest that the mechanism is inactive for the considered strains of *S. aureus* or the enzyme production rates are too small, which we are not able to account by our method. Using our approach we identified (by the highest support) the gene *ermA* responsible for the most common drug resistance mechanism.

Here, there is one erythromycin-susceptible strain, *USA300 TCH959*, which harbours the *ermA* gene. This may suggest disruption of the drug resistance mechanism in that strain, errors in drug susceptibility testing or errors in sequencing.

Interestingly, we identified gene *SAR1736(spc2)* (which is a known spectinomycin resistance determinant) as potentially associated with erythromycin resistance. This suggests that drug resistance to spectinomycin and erythromycin co-evolved, despite these two drugs belonging to different classes according to the ATC drug classification system (WHO, 2010).

Gentamicin Gentamicin works by inhibition of protein synthesis by binding the 30S subunit of the ribosome complex (Shakil et al., 2008).

Interestingly, strain *USA300 FPR3757* exhibits intermediate drug resistance, which is correlated with the absence of *aacA-aphD* gene in its genome sequence. Since our method requires binary information on drug susceptibility, we marked this strain as drug-susceptible for experiments.

The most common resistance mechanism responsible for high levels of gentamicin resistance is mediated by the drug-modifying enzyme *SaurJH1 2806(aacA-aphD)*. Applying our methodology we identified the gene encoding it as likely to be associated with drug resistance (maximal support). Moreover, we identified also the gene *SaurJH1 2805* as putatively associated with gentamicin resistance. The close proximity of these two genes in the genomes and their highly correlated gene-gain/-loss profiles suggest co-evolution. We hypothesize that the gene *SaurJH1 2805* plays some role in drug resistance for gentamicin.

1.3.2 M. tuberculosis datasets

Here we present results obtained by applying GWAMAR to two large datasets for *M. tuberculosis*. We use these case studies to present the usability of GWAMAR to identify chromosomal mutations associated with drug resistance. The first dataset is prepared for the set of 173 strains with genome sequences and annotations publicly available in the PATRIC database, developed by Wattam et al. (2014). For this set of strains, we collected drug resistance information from over 20 publications. The genotype and phenotype data for the second dataset comes from the *M. tuberculosis* Drug Resistance Directed Sequencing Database at `http://www.broadinstitute.org/annotation/genome/mtb_drug_resistance`.

1.3.2.1 First case study

The first case study is based on the set of 173 fully sequenced strains of *M. tuberculosis* with publicly available data.

The preprocessing steps of preparing the genotype data were performed using eCAMBer, our tool to support comparative analysis of multiple bacterial strains (Woźniak et al., 2012).

In particular, first, we used eCAMBer to download the genome sequences and annotations from the PATRIC database (Wattam et al., 2014). Next, we applied eCAMBer to unify the genome annotations of protein-coding genes and to identify the clusters of genes with high sequence similarity. Then, for the subset of 4379 such identified gene clusters with genes present in at least 90% of the strains, we computed multiple alignments using MUSCLE (Edgar, 2004). The multiple alignments were computed for amino acid sequences of protein coding genes, as well as nucleotide sequences of their promoter regions (-50bp upstream), and rRNA genes. In total, based on the computed multiple alignments, we identified 118913 mutations, which constituted the input genotype data for GWAMAR. After the procedure of binarization in GWAMAR we ended up with 18635 binary mutation profiles.

The input phenotype data was collected from over 20 publications issued together with the fully sequenced genomes. Based on the drug resistance information collected for ciprofloxacin and ofloxacin, we introduced a new drug resistance profile for the drug family of fluoroquinolones. A strain was categorized as susceptible to fluoroquinolones if it was susceptible to at least one of the drugs, but not resistant to any of them. Similarly, a strain was categorized as resistant to fluoroquinolones if it was resistant to at least one of the drugs, but not susceptible to any of them. If none of the cases was satisfied for a strain, then the drug resistance status of the strain was categorized as unknown. We restrict analysis to the set of six drugs or drug families: fluoroquinolones, ethambutol, isoniazid, pyrazinamide, rifampicin and streptomycin.

The input phylogenetic tree was reconstructed using the maximum likelihood approach implemented in the PhyML, developed by (Guindon et al., 2010). As for the input for the tool we used the set of all the identified point mutations concatenated into one multiple alignment file. This input was prepared by an additional feature of eCAMBer.

Having prepared the set of binary mutation profiles and drug resistance profiles, together with the phylogenetic tree, we applied GWAMAR to compute MI, OR, H, WS, TGH and RBM association scores. However, in order to compute TGH score efficiently, we averaged three TGH scores obtained over three random binary trees we obtained from the original tree by splitting its nodes with multifurcations. This step of computations, ran using 6 processors, took around 6s for MI, OR, H and WS; around 34s for TGH; and around 3s for all the considered variants of the RBM score.

drug name	**gene id**	**gene name**	**mutation**	**all**	**h.c.**	**TGH**
Fluoroquinolones	Rv0006	gyrA	$D94H_1A_5N_2Y_2G_{12}$	Y	Y	14.184
Isoniazid	Rv1908c	katG	$S315N_1G_2T_{75}$	Y	Y	9.045
Rifampicin	Rv0667	rpoB	$S450L_{71}$	Y	Y	8.602
Streptomycin	Rv0682	rpsL	$K43R_{15}$	Y	Y	8.323
Ethambutol	Rv3795	embB	$M306L_1I_{32}V_{18}$	Y	Y	8.250
Isoniazid	Rv1483	fabG1	$C\text{-}15T_{30}$	Y	Y	5.845
Rifampicin	Rv0667	rpoB	$D435Y_2F_5V_{11}G_3A_1$	Y	Y	5.040
Streptomycin	Rv0682	rpsL	$K88R_5M_1$	Y	Y	4.164
Ethambutol	Rv3795	embB	$E504G_1D_1$	N	N	3.331
Pyrazinamide	Rv2043c	pncA	$H51P_1$	Y	Y	2.708
Pyrazinamide	Rv2043c	pncA	$W68L_1$	Y	Y	2.708
Rifampicin	Rv0667	rpoB	$H445D_8Y_2R_1$	Y	Y	2.530
Streptomycin	Rvnr01	rrs	$G1108C_2$	N	N	1.717
Ethambutol	Rv3795	embB	$D869G_1$	N	N	1.688
Ethambutol	Rv3795	embB	$A505T_1$	N	N	1.688
Ethambutol	Rv3795	embB	$D1024N_1$	Y	N	1.688
Fluoroquinolones	Rv0005	gyrB	$N538T_1$	Y	Y	1.685
Fluoroquinolones	Rv0006	gyrA	$S91P_1$	Y	Y	1.685
Fluoroquinolones	Rv0005	gyrB	$T539I_1$	N	N	1.685
Streptomycin	Rvnr01	rrs	$A1401G_{17}$	Y	N	1.288
Ethambutol	Rv3795	embB	$Y334H_2$	Y	N	1.054
Ethambutol	Rv3795	embB	$Q497R_2$	Y	Y	1.054
Rifampicin	Rv0667	rpoB	$E250G_3$	N	N	1.047
Fluoroquinolones	Rv0006	gyrA	$A90V_6G_3$	Y	Y	1.035
Streptomycin	Rvnr01	rrs	$C517T_{33}$	Y	Y	0.915

Table 1.4: 25 top-scoring associations between drug resistance profiles and point mutations in the case study on 173 fully sequenced *M. tuberculosis* strains, when restricted to only these genes which are associated with drug resistance to the corresponding drugs

. Each row corresponds to one association, whereas the consecutive columns describe: drug name, gene identifier, gene name, mutation, association presence in the TBDReaMDB database, status indicating whether the association is categorized as high-confidence in TBDReaMDB, TGH score. Lower indexes in the mutation descriptions indicate the numbers of strains possessing the corresponding amino acid or nucleotide variant.

We took a closer look at the top-scoring mutations returned by the scores, but restricting our analysis to only these associations which involve genes which are known to be associated with drug resistance for the corresponding drug — possessing at least one point mutation annotated as high-confidence in the TBDReaMDB database. Table 1.4 presents the list of top 25 associations ordered according to TGH score. In the set of 25 top-scored associations, 19 are present in the TBDReaMDB database and 16 of them are categorized as high-confidence mutations. A closer look at the mutations which are not present in TBDReaMDB revealed that some of them can be supported by literature. In particular, mutation $E504G/D$ in *embB* has recently been reported as associated with resistance to ethambutol (Bakuła et al., 2013). The close proximity of this mutation to $A505T$ in *embB* may also suggest that the latter is associated with ethambutol resistance. Similarly, the mutation $T539I$ has already been associated with resistance to fluoroquinolones (Malik et al., 2012).

Literature search did not provide us any additional support for the remaining three mutations ($D869G$ in *embB* and $G1108C$ in *rrs*), which haven't been also reported in TBDReaMDB.

1.3.2.2 Second case study

The second case study, *mtu_broad*, is based on the data available in the Broad Institute database http://www.broadinstitute.org/annotation/genome/mtb_drug_resistance. This database contains sequencing data and drug resistance information for 1398 strains of *M. tuberculosis*. However, it should be noted that only genes of interest were sequenced; Table 1.5 presents the list of 28 sequenced genes for each strain. Additionally 12 promoter sequences were sequenced. In total, this database contains 1067 mutations (non-synonymous amino-acid changes or nucleotide changes in promoters), which constituted the input genotype data for GWAMAR. After the procedure of binarization in GWAMAR we ended up 850 binary mutation profiles.

Similar to the previous case study, based on the drug resistance information available in the database for ciprofloxacin, ofloxacin, levofloxacin and moxifloxacin, we introduced a new drug resistance profile for the drug family of fluoroquinolones. A strain was categorized as susceptible to fluoroquinolones if it was susceptible to at least one of the drugs, but not resistant to any of them. Similarly, a strain was categorized as resistant to fluoroquinolones if it was resistant to at least one of the drugs, but not susceptible to any of them. If none of the cases was satisfied for a strain, then the drug resistance status of the strain was categorized as unknown. We restrict further analysis to the set of six drugs or drug families: fluoroquinolones, ethambutol, isoniazid, pyrazinamide, rifampicin and streptomycin.

In these experiments the phylogenetic tree was reconstructed using the maximum likelihood approach implemented in the PhyML package, developed by Guindon et al. (2010). As an input for the tool we used the set of all available mutations

gene id	gene name	description	prom. sequenced?
Rv0005	gyrB	DNA gyrase subunit B	yes
Rv0006	gyrA	DNA gyrase subunit A	yes
Rv0341	iniB	isoniazid inductible gene protein	yes
Rv0342	iniA	isoniazid inductible gene protein	yes
Rv0343	iniC	isoniazid inductible gene protein	yes
Rv0667	rpoB	DNA-directed RNA polymerase beta chain	yes
Rv0682	rpsL	30S ribosomal protein S12	yes
Rv1483	fabG1	3-oxoacyl-[acyl-carrier protein] reductase	yes
Rv1484	inhA	NADH-dependent enoyl-[acyl-carrier-protein] reductase	yes
Rv1694	tlyA	cytotoxin|haemolysin	no
Rv1854c	ndh	NADH dehydrogenase	yes
Rv1908c	katG	catalase-peroxidase-peroxynitritase T	no
Rv2043c	pncA	pyrazinamidase/nicotinamidas	yes
Rv2245	kasA	3-oxoacyl-[acyl-carrier protein] synthase 1	no
Rv2427Ac	oxyR'	hypothetical protein	no
Rv2428	ahpC	alkyl hydroperoxide reductase C protein	yes
Rv2764c	thyA	thymidylate synthase	yes
Rv2764c	ddl	D-alanine-D-alanine ligase ddlA	no
Rv3423c	alr	alanine racemase	no
Rv3793	embC	membrane indolylacetylinositol arabinosyltransferase	yes
Rv3794	embA	membrane indolylacetylinositol arabinosyltransferase	yes
Rv3795	embB	membrane indolylacetylinositol arabinosyltransferase	yes
Rv3854c	ethA	monooxygenase	yes
Rv3919c	gid	glucose-inhibited division protein B	yes
Rvnr01	rrs	ribosomal RNA 16S	no
Rvnr02	rrl	ribosomal RNA 23S	no

Table 1.5: List of sequenced genes and promoters available at the Broad Institute database, `http://www.broadinstitute.org/annotation/genome/mtb_drug_resistance`.

concatenated into one multiple alignment file. The preparation of the multiple alignment file as well as running PhyML was done with the use of eCAMBer.

Similarly, as in the *mtu173* dataset, we applied GWAMAR to compute MI, OR, H, WS, TGH and RBM association scores. As in the previously described case study, in order to compute TGH score efficiently, we averaged three TGH scores obtained over three random binary trees we obtained from the original tree by splitting its nodes with multifurcations. This step of computations, ran using 6 processors, took around 5s for MI, OR, H and WS; around 2h for TGH; and around 2s for all the considered variants of the RBM score. It took relatively long time to compute TGH score due to its high time complexity with respect to the numbers of strains considered 1.2.3.

Similarly as for the *mtu173* dataset, we sort the set of putative associations according to the TGH score, but restricting our analysis to only these associations which involve genes which are known to be associated with drug resistance for the corresponding drug — possessing at least one point mutation annotated as high-confidence in the TBDReaMDB database. Table 1.6 presents the list of the

drug name	**gene id**	**gene name**	**mutation**	**all**	**h.c.**	**TGH**
Fluoroquinolones	Rv0006	gyrA	$D94Y_6H_2A_{26}G_{78}N_{14}$	Y	Y	128.323
Rifampicin	Rv0667	rpoB	$S450L_{743}W_{22}$	Y	Y	72.284
Ethambutol	Rv3795	embB	$M306T_1L_{16}V_{290}I_{313}$	Y	Y	70.217
Fluoroquinolones	Rv0006	gyrA	$A90G_2V_{46}$	Y	Y	41.699
Streptomycin	Rv0682	rpsL	$K43R_{228}$	Y	Y	30.012
Isoniazid	Rv1908c	katG	$S315T_{895}G_2I_3R_3N_{27}$	Y	Y	27.966
Ethambutol	Rv3795	embB	$Q497H_5K_{18}P_{10}R_{43}$	Y	Y	17.081
Streptomycin	Rv0682	rpsL	$K88Q_1R_{28}T_{32}M_7$	Y	Y	16.327
Fluoroquinolones	Rv0005	gyrB	$N538K_1S_1T_9D_2$	Y	Y	12.605
Rifampicin	Rv0667	rpoB	$H445P_2Q_2L_{27}Y_{53}R_{42}D_{25}N_7$	Y	Y	12.252
Streptomycin	Rvnr01	rrs	$A1401G_{254}$	Y	N	9.509
Streptomycin	Rvnr01	rrs	$A514C_{90}$	Y	Y	8.940
Pyrazinamide	Rv2043c	pncA	$T135A_1P_{22}$	Y	N	8.814
Fluoroquinolones	Rv0006	gyrA	$S91P_9$	Y	Y	7.557
Rifampicin	Rv0667	rpoB	$D435H_1N_2A_2Y_{27}G_3V_{140}$	Y	Y	7.480
Ethambutol	Rv3795	embB	$G406C_3A_{68}D_{52}S_{43}$	Y	Y	7.057
Pyrazinamide	Rv2043c	pncA	$T\text{-}11G_3C_{24}$	Y	Y	6.766
Fluoroquinolones	Rv0006	gyrA	$D89G_2N_4$	Y	N	6.253
Pyrazinamide	Rv2043c	pncA	$L120P_{20}R_5$	Y	N	6.146
Streptomycin	Rvnr01	rrs	$C517T_{26}$	Y	Y	5.169
Pyrazinamide	Rv2043c	pncA	$Q10H_3R_{10}P_{12}$	Y	Y	5.053
Pyrazinamide	Rv2043c	pncA	$V139M_3G_2A_7L_1$	Y	Y	5.053
Ethambutol	Rv3795	embB	$D328G_5H_1Y_9$	Y	N	5.032
Streptomycin	Rvnr01	rrs	$A908C_7G_1$	Y	N	4.779
Pyrazinamide	Rv2043c	pncA	$D12E_1G_5N_1A_{12}$	Y	Y	4.725

Table 1.6: 25 top-scored associations between drug resistance profiles and point mutations in the case study for 1398 partially sequenced *M. tuberculosis* strains, when restricted to only these genes which are associated with drug resistance to the corresponding drugs. This dataset is provided by The Broad Institute. Each row corresponds to one association, whereas the consecutive columns describe: drug name, gene identifier, gene name, description of the mutation, association presence in the TBDReaMDB database, status indicating whether the association is categorized as high-confidence in TBDReaMDB, TGH score. Lower indexes in the mutation descriptions indicate the numbers of strains possessing the corresponding amino acid or nucleotide variant.

top 25 associations ordered according to TGH score. In the set of 25 top-scored associations, all are present in TBDReaMDB and 19 of them are categorized as high-confidence mutations. The presence in the TBDReaMDB database provides some additional support for the six associations which are categorized as high-confidence.

1.3.2.3 Assessment of accuracy

Here we use the two datasets described above to assess the accuracy of the various association scores, viz: mutual information, odds ratio, hypergeometric, weighted support, TGH and RBM.

We considered to use for comparison CCTSWEEP, proposed by Habib et al.

(2007) — a score conceptually similar to the TGH score. However, we failed to run its implementation, probably, due to rather poor documentation. Its authors had not responded to our queries in time. Thus, we omitted it from our experiments.

In order to assess the accuracy of different association scores we need a reliable dataset of known drug resistance associations. Here, we test two approaches to define our gold standard. In the first, we take all 607 associations from the Tuberculosis Drug Resistance Mutation Database (TBDReaMDB), developed by Sandgren et al. (2009). In the second, we use the subset of 81 drug resistance associations classified as *high-confidence* in the database. Table 1.6 presents the list of the mutations in TBDReaMDB with the distinguished subset of high-confidence associations. In all comparative experiments we assume a putative association to be a positive if it is present in the gold standard.

In both case studies, as the set of positives, we assume the subset of mutations present in our gold standard, also present in the available genotype data. This is the set of mutations which may be potentially detected (we say they are "detectable") using the available datasets. Thus, in the case when all TBDReaMDB associations constitute the gold standard, there are 94 and 212 of such "detectable" associations for the *mtu173* and *mtu_broad* datasets, respectively. Likewise, if only high-confidence associations are considered as gold standard, then 39 and 74 of such "detectable" associations for the *mtu173* and *mtu_broad* datasets, respectively

The set of negatives is constructed by random sampling from the whole set of identified putative associations except for the associations which are classified as positives. The number of sampled negatives equals the total number of mutations present in the genes which have at least one mutation in the gold standard. It should be noted that, among the mutations present within the genes which are associated with drug resistance, many can be real positives (associated with drug resistance), but lacking the annotation in TBDReaMDB. Thus, we use this approach of sampling from all mutations in order to significantly reduce the probability of classifying as negatives associations which are real but not present in the database.

Figure 1.7 presents statistics for the Area Under the Curve (AUC) for the precision and recall curves for different association scores. The results come from

drug name	**gene id**	**gene name**	**positions**	
			TBDReaMDB (high-conf.)	**TBDReaMDB (all)**
Ethambutol	Rv0340			173
	Rv0341	iniB		-89,47
	Rv0342	iniA		308,501
	Rv0343	iniC		248,351
	Rv1267c	embR		7,32,53... (24 in total)
	Rv3124	moaR1		-16
	Rv3125			54
	Rv3126			276
	Rv3264c	manB		152
	Rv3266c	rmlD		-71,257,284
	Rv3793	embC		5,244,247... (25 in total)
	Rv3794	embA		-16,-12,-11... (10 in total)
	Rv3795	embB	306,406,497	128,221,225... (85 in total)
Fluoroquinolones	Rv0005	gyrB	538	457,458,472... (9 in total)
	Rv0006	gyrA	90,91,94,102,126	74,80,88... (10 in total)
Isoniazid	Rv0129c	fbpC		-63,-23,158
	Rv0340			163
	Rv0342	iniA		3,537
	Rv0343	iniC		83
	Rv1483	fabG1	-15,-8	-92,-67,-24... (10 in total)
	Rv1484	inhA		8,16,21... (10 in total)
	Rv1592c			-29,42,430
	Rv1772			4
	Rv1854c	ndh		13,18,110,239,268
	Rv1908c	katG	279,315	1,2,11... (171 in total)
	Rv1909c	furA		5
	Rv2243	fabD		275
	Rv2245	kasA	269	66,77,121... (7 in total)
	Rv2247	accD6		229
	Rv2428	ahpC	-46,-39,21	-66,-49,-46... (21 in total)
	Rv2846c	efpA		73
	Rv3566c	nat		67,207
	Rv3795	embB		333
Pyrazinamide	Rv2043c	pncA	-11,7,10... (51 in total)	-12,-11,-7... (103 in total)
Rifampicin	Rv0667	rpoB	432,435,441,445,450,452	65,300,409... (38 in total)
Streptomycin	Rv0682	rpsL	43,88	9,40,41... (11 in total)
	Rv3919c	gidB		16,40,45... (19 in total)
	Rvnr01	rrs	492,513,514,517,907	190,427,462... (17 in total)

Figure 1.6: The list of drug resistance associations in the TBDReaMDB database. The first three columns correspond to the drug name, gene identifier and gene name of the gene corresponding to the point mutation. The next column lists the positions of the mutations corresponding to associations classified as *high-confidence* in the TBDReaMDB database. The last column lists positions of all the mutations present in the database. Each positive number indicates the position of the mutation in the amino acid sequence of the corresponding gene. Each negative number indicates the position of the mutation in the nucleotide sequence of the promoter of the gene, counting upstream its TIS. In some cases (if there are too many mutation to fit them within the table width) we do not list them all here — the complete list might be accessed at the project website, http://bioputer.mimuw.edu.pl/gwamar.

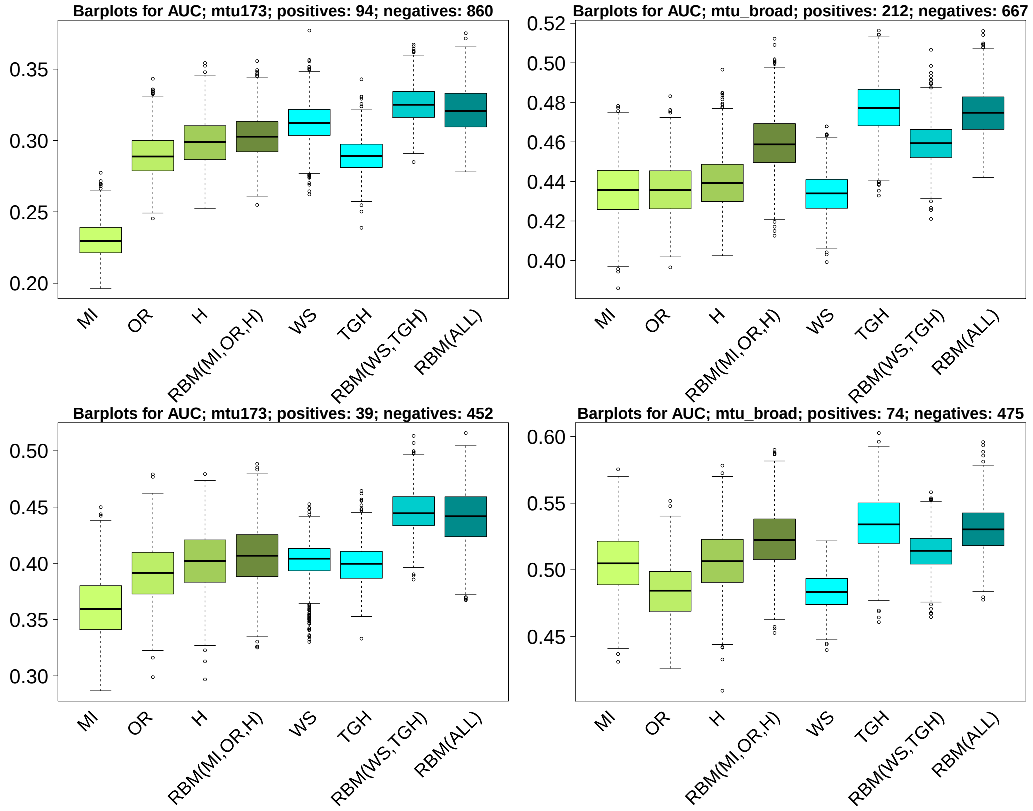


Figure 1.7: Comparison of different association scores implemented in GWAMAR based on the Area Under the Curve (AUC) statistic for the precision-recall curves. Left panels present the results for the *mtu173* dataset; right for the *mtu_broad* dataset. The first row of panels corresponds to the experiments in which all associations present in TBDReaMDB were used as the gold standard, whereas the second row corresponds to the experiments in which only high-confidence associations were used as the gold standard. The process of sampling the set of negatives was repeated 1000 times. The barplots for tree-ignorant and tree-aware scores are shown green and blue, respectively.

sampling the set of negatives and calculating the AUC, repeated 1000 times. The results show that on average, the tree-aware statistics (WS, TGH) performed slightly better than the tree ignorant scores on the *mtu173* dataset. They also show that, TGH performed best on the large *mtu_broad* dataset, but was slightly worse on the relatively small *mtu173* dataset. The presented results also show, that the Rank-based metascore performs consistently better than individual scores in most of the settings. For example, the RBM (MI,OR,H) outperformed all individual scores it combines in all the settings. Notably, the RBM(ALL) score, outperformed consistently all tree-ignorant scores in all the settings.

We conclude that, the tree-aware association scores outperform the tree-ignorant methods. In particular, the Rank-based metascore performed consistently better than the individual scores. However, the advantage is rather small and dependent on a setting. The performance may be influenced by the tree topology, the strains number or the small number of positives.

1.3.2.4 Compensatory mutations

The most common mechanism of rifampicin resistance in *M. tuberculosis* is acquired by point mutations within the rifampicin resistance determining region (RRDR) in the *rpoB* gene, which corresponds to the rifampicin binding spot (Patra et al., 2010).

Since the *rpoB* gene is essential for bacteria, mutations present in this gene, due to altering its structure, have often deleterious effect on the bacterial fitness in the drug-free environment (Brandis and Hughes, 2013). This effect may be potentially reversed by compensatory mutations. Thus, compensatory mutations tend to appear later, in the evolutionary history, than the mutations directly responsible for drug resistance. Hence, for a given compensatory mutation, we expect to observe it in a subset of strains which correspond to the mutation directly responsible for drug resistance.

Based on the above described assumption, we identify putative compensatory mutations using the following procedure applied to the *mtu_broad* and *mtu173* datasets. First, we identify the set of mutations within RRDR. Here, RRDR is defined as a region of 27 amino acids between positions 426 and 452 in the *rpoB* gene. Mutations from this region constitute the set of putative primary (directly responsible for drug resistance) mutations. For the *mtu173* dataset we obtained the following list of putative primary mutations:

- $L430P_1$
- $D435Y_2F_5V_{11}G_3A_1$
- $H445D_8Y_2R_1$
- $S450L_{71}$
- $L452P_2$

Here, the description of each mutation comprises of the reference amino acid, the position of the mutation in the gene, and different amino-acid variants of the mutation among the strains. For each mutation, the lower indexes indicate the number of strains possessing the corresponding amino-acid variant of the mutation within the 173 strains in the *mtu173* dataset.

Applying the same method to the *mtu_broad* dataset we obtained the following list of putative primary mutations:

- $S428R_2$
- $Q429P_1H_1$
- $L430P_3R_9$
- $S431G_1$
- $Q432P_5E_2L_1K_5H_5$
- $M434I_2$
- $D435H_1N_2A_2Y_{27}G_3V_{140}$
- $N437H_1K_1$
- $N438H_1$
- $P439S_1$
- $S441L_4$
- $H445P_2Q_2L_{27}Y_{53}R_{42}D_{25}N_7$
- $R448Q_7$
- $S450L_{743}W_{22}$
- $L452V_1P_{24}$

Here, similarly, as for the previously described list of mutations, for each mutation, the lower indexes indicate the number of strains possessing the corresponding amino-acid variant of the mutation within the 173 strains in the *mtu_broad* dataset.

Interestingly, in both case studies the sets of strains possessing the primary mutations tend to be disjoint. For example, for the *mtu173* dataset, the sets of strains possessing mutations at positions 450 and 435 are disjoint (hypergeometric test p-value=$2.302 \cdot 10^{-8}$). The sets of strains possessing the mutations at positions 450 and 445 are also disjoint (hypergeometric test p-value=0.00026). Similarly, for the *mtu_broad* dataset, the sets of strains possessing mutations at positions 450 and 445 are disjoint (hypergeometric test p-value=$2.62 \cdot 10^{-63}$). The sets of strains possessing mutations at positions 450 and 435 overlap by only three elements (hypergeometric test p-value=$3.87 \cdot 10^{-64}$). We hypothesize that this phenomenon may be caused by the negative epistatic interactions between mutations from RRDR (Khan et al., 2011).

Finally, we identify a set of putative compensatory mutations, applying the following simple rule: a mutation is classified as a putative compensatory mutation if the set of strains possessing the mutation is contained within the set of strains corresponding to one of the mutations identified as primary mutations (from RRDR).

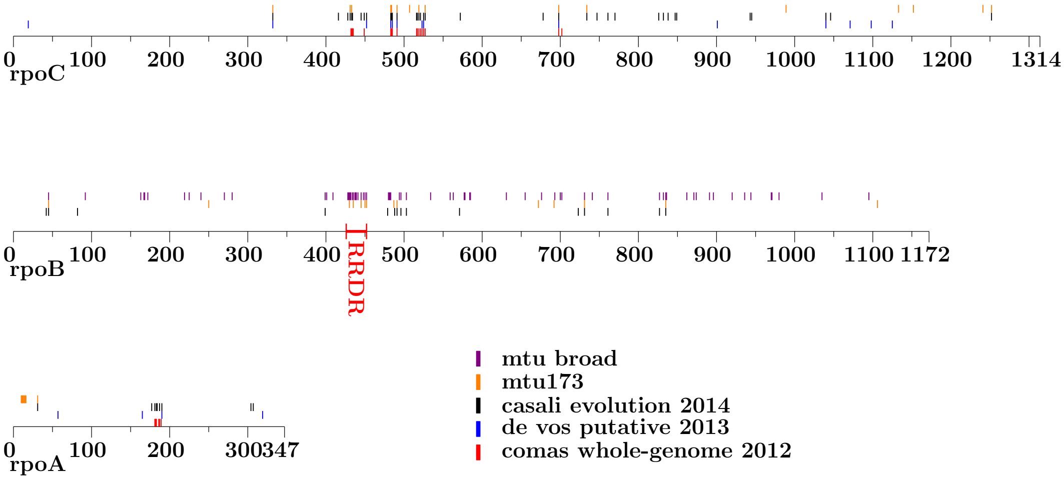


Figure 1.8: Comparison of the sets of putative compensatory mutations within the *rpoA*, *rpoB* and *rpoC* genes, reported in various sources and detected in our two datasets. Each mutation's position is indicated by a vertical line of the color corresponding to the source it was reported in. In particular orange and violet lines indicate positions of mutations identified by our approach applied to the *mtu173* and *mtu_broad* datasets, respectively. The other lines indicate mutations reported in the recent articles by Comas et al. (2012) (red), de Vos et al. (2013) (blue) and Casali et al. (2014) (black).

Here, we compare the sets of putative compensatory mutations for rifampicin within the *rpoA*, *rpoB* and *rpoC* genes, identified by our approach with the mutations reported in other recent articles by Comas et al. (2012), de Vos et al. (2013), and Casali et al. (2014).

Figure 1.8 presents the distribution of the putative compensatory mutations identified in these recent studies, put together with the set of putative compensatory mutations identified based on our two case studies. Note that the identified putative compensatory mutations tend to cluster within the region of the *rpoC* gene from 430 to 530.

Table 1.7 presents another view on the list of putative compensatory mutations identified by our approach with comparison with those reported in the other recent articles. However, due to space limitation, in this table, we only list a subset of such mutations. A mutation is listed in the table, if it was identified in one of our two datasets and also reported in at least one of the three recent articles, or reported in at least two of the articles. The complete list of putative compensatory mutations is available on the website of our project. `http://bioputer.mimuw.edu.pl/gwamar`.

We conclude that using our approach we were able to re-identify most of the putative compensatory mutations identified previously. Moreover, in contrast to the other research groups, which used in house sequenced bacteria, we achieved our results by analysis on freely and publicly available data.

1.4 Summary

In this chapter, we presented GWAMAR, a tool we have developed for identifying of drug resistance-associated mutations based on comparative analysis of whole-genome sequences in bacterial strains.

The tool is designed as an automatic pipeline which employs eCAMBer for preprocessing of the genotype data. This preprocessing includes: (i) downloading of genome sequences and gene annotations, (ii) unification of gene annotations among the set of considered strains, (iii) identification of gene families, (iv) computation of multiple alignments and identification of point mutations which constitute the input genotype data.

GWAMAR implements various statistical methods — such as mutual information, odds ratio, hypergeometric score — to associate the drug resistance phe-

gene	position	ref aa	comas 2012	de vos 2013	casali 2014	mtu173	mtu broad
rpoA	31	G			A_1S_1	A_1S_1	
rpoA	181	T	I_1		A_1		
rpoA	183	V	A_1G_1		G_1		
rpoA	187	T	A_3P_1		A_1P_1		
rpoA	190	D		G_1	G_1		
rpoB	45	P			S_1	L_1	$A_3L_9S_7T_2$
rpoB	399	T			A_1		A_5I_4
rpoB	491	I			V_1	L_1	
rpoB	496	V			L_1M_1		G_1M_3
rpoB	503	F			S_1		S_3
rpoB	731	L			P_1	P_1	P_8
rpoB	761	E			D_1		D_1
rpoB	827	R			C_1		C_3
rpoB	835	H			P_1R_1	R_1	P_1R_3
rpoC	332	G		R_2	$C_1R_1S_1$	S_7	
rpoC	431	V			M_1	M_1	
rpoC	433	G	S_1		C_1S_1	S_1	
rpoC	434	P	A_1R_1		Q_1R_1		
rpoC	449	L	V_1		V_1		
rpoC	452	F		C_1	C_1		
rpoC	483	V	A_3G_3	A_1G_3	A_1G_1	A_1G_5	
rpoC	484	W	G_2		G_1	G_1	
rpoC	485	D	H_1N_1	Y_1	N_1Y_1		
rpoC	491	I	T_1V_2	T_2	T_1V_1	T_2	
rpoC	516	L	P_2		P_1		
rpoC	519	G	D_1		D_1	V_1	
rpoC	521	A	D_1		D_1		
rpoC	525	H	N_1	Q_1	Q_1		
rpoC	527	L	V_1		V_1	V_8	
rpoC	698	N	$H_1K_1S_1$	H_1S_1	$H_1K_1S_1$	K_1	
rpoC	734	A			V_1	V_2	
rpoC	1040	P		R_1	$R_1S_1T_1$		
rpoC	1252	V			L_1	M_4	

Table 1.7: The list of putative compensatory mutations identified by our approach applied to the *mtu_broad* and *mtu173* datasets, identified in one of our two datasets and also reported in at least one of the three recent articles, or reported in at least two of the articles. The first two columns correspond to the gene name, and the reference amino acid, respectively. The next three columns provide brief descriptions of the mutations identified in the three recent studies: by Comas et al. (2012), de Vos et al. (2013) and Casali et al. (2014), respectively. The last two columns list the mutations identified based on our two case studies. Each mutation's description comprises of the reference amino acid, the position of the mutation in the gene, and different amino-acid variants of the mutation among the strains. For each mutation, the lower indexes indicate the number of strains, in the corresponding dataset, possessing the corresponding amino-acid variant of the mutation.

notypes with point mutations. In this work, we also present weighted support (WS) and tree-generalized hypergeometric (TGH) score — two new statistical scores — which employ phylogenetic information. As a part of this work, we also present yet another score, called Rank-based metascore (RBM) to combine multiple scores, thus compromising for weak points of the individual scores being combined.

In order to test our approach, we prepared one dataset for *S. aureus* and two datasets for *M. tuberculosis*. The presented results demonstrate usefulness of our approach to identify drug-resistance associated mutations based on publicly available sequencing data. In particular, we were able to re-identify most of the known drug-resistance associations. Our results also support the phenomena previously reported in the literature, such as: (i) drug resistance-associated mutations tend to have multiple variants observed; or (ii) drug resistance associated mutations tend to cluster together in close genomic proximity.

Moreover, since most of the recent studies on the subject of compensatory mutations and in general drug resistance-associated mutations used in-house sequenced bacteria, we achieved our promising results basing our analysis solely on freely available public data.

The presented results also suggest that tree-aware methods (WS and TGH) perform better than methods which do not incorporate phylogenetic information. The results also show that the RBM score outperforms the individual scores in most of the settings. In particular, the RBM (ALL) score performed better than any tree-ignorant score in all the experiments.

Finally, despite some promising results, the presented tool has some limitations. First, it does not take into account epistatic interactions between mutations. Second, it takes into account only genomic changes, ignoring levels of gene expression. Thirdly, it provides putative *in silico* associations which should be subjected to further investigation in wet lab experiments.

The tools, case-study input data and the obtained results are available at the website of this project, http://bioputer.mimuw.edu.pl/gwamar.

References

Bakuła, A, Z., Napiórkowska, Rkowska, A., Bielecki, J., Augustynowicz-Kopeć, Ewa, Zwolska, Z., Jagielski, T., 2013. Mutations in the embB gene and their association with ethambutol resistance in multidrug-resistant Mycobacterium tuberculosis clinical isolates from poland, BioMed Research International, 2013, p. e167954.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Sayers, E. W., 2013. GenBank, Nucleic acids research, 41(Database issue), p. D36.

Boto, L., 2010. Horizontal gene transfer in evolution: facts and challenges, Proceedings of the Royal Society of London B: Biological Sciences, 277(1683), p. 819.

Brandis, G., Hughes, D., 2013. Genetic characterization of compensatory evolution in strains carrying rpoB ser531leu, the rifampicin resistance mutation most frequently found in clinical isolates, Journal of Antimicrobial Chemotherapy, 68(11), p. 2493.

Cabrera, C. P., Navarro, P., Huffman, J. E., Wright, A. F., Hayward, C., Campbell, H., Wilson, J. F., Rudan, I., Hastie, N. D., Vitart, V., Haley, C. S., 2012. Uncovering networks from genome-wide association studies via circular genomic permutation, G3 (Bethesda, Md.), 2(9), p. 1067.

Casali, N., Nikolayevskyy, V., Balabanova, Y., Harris, S. R., Ignatyeva, O., Kontsevaya, I., Corander, J., Bryant, J., Parkhill, J., Nejentsev, S., Horstmann, R. D., Brown, T., Drobniewski, F., 2014. Evolution and transmission of drug-resistant tuberculosis in a russian population, Nature Genetics, 46(3), p. 279.

Chopra, I., Roberts, M., 2001. Tetracycline antibiotics: mode of action, applications, molecular biology, and epidemiology of bacterial resistance, Microbiology and Molecular Biology Reviews: MMBR, 65(2), p. 232.

Clarke, G. M., Anderson, C. A., Pettersson, F. H., Cardon, L. R., Morris, A. P., Zondervan, K. T., 2011. Basic statistical analysis in genetic case-control studies, Nature protocols, 6(2), p. 121.

Comas, I., Borrell, S., Roetzer, A., Rose, G., Malla, B., Kato-Maeda, M., Galagan, J., Niemann, S., Gagneux, S., 2012. Whole-genome sequencing of rifampicin-resistant mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes, Nature Genetics, 44(1), p. 106.

Connell, S. R., Tracz, D. M., Nierhaus, K. H., Taylor, D. E., 2003. Ribosomal potection proteins and their mechanism of tetracycline resistance, Antimicrobial Agents and Chemotherapy, 47(12), p. 3675.

Daubin, V., Moran, N. A., Ochman, H., 2003. Phylogenetics and the cohesion of bacterial genomes, Science, 301(5634), p. 829.

Davies, G., Tenesa, A., Payton, A., Yang, J., Harris, S. E., Liewald, D., Ke, X., Le Hellard, S., Christoforou, A., Luciano, M., McGhee, K., Lopez, L., Gow, A. J., Corley, J., Redmond, P., Fox, H. C., Haggarty, P., Whalley, L. J., McNeill, G., Goddard, M. E., Espeseth, T., Lundervold, A. J., Reinvang, I., Pickles, A., Steen, V. M., Ollier, W., Porteous, D. J., Horan, M., Starr, J. M., Pendleton, N., Visscher, P. M., Deary, I. J., 2011. Genome-wide association studies establish that human intelligence is highly heritable and polygenic, Molecular Psychiatry, 16(10), p. 996.

de Vos, M., Muller, B., Borrell, S., Black, P. A., van Helden, P. D., Warren, R. M., Gagneux, S., Victor, T. C., 2013. Putative compensatory mutations in the rpoC gene of rifampin-resistant mycobacterium tuberculosis are associated with ongoing transmission, Antimicrobial Agents and Chemotherapy, 57(2), p. 827.

Drawz, S. M., Bonomo, R. A., 2010. Three decades of beta-lactamase inhibitors, Clinical Microbiology Reviews, 23(1), p. 160.

Dutheil, J. Y., 2012. Detecting coevolving positions in a molecule: why and how to account for phylogeny, Briefings in Bioinformatics, 13(2), p. 228.

Edgar, R. C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Research, 32(5), p. 1792.

Farhat, M. R., Shapiro, B. J., Kieser, K. J., Sultana, R., Jacobson, K. R., Victor, T. C., Warren, R. M., Streicher, E. M., Calver, A., Sloutsky, A., Kaur, D., Posey, J. E., Plikaytis, B., Oggioni, M. R., Gardy, J. L., Johnston, J. C., Rodrigues, M., Tang, P. K. C., Kato-Maeda, M., Borowsky, M. L., Muddukrishna, B., Kreiswirth, B. N., Kurepina, N., Galagan, J., Gagneux, S., Birren, B., Rubin, E. J., Lander, E. S., Sabeti, P. C., Murray, M., 2013. Genomic analysis identifies targets of convergent positive selection in drug-resistant mycobacterium tuberculosis, Nature Genetics, 45(10), p. 1183.

Felsenstein, J., 2005. PHYLIP (Phylogeny Inference Package) version 3.6, Department of Genome Sciences, University of Washington, Seattle.

Ferrero, L., Cameron, B., Crouzet, J., 1995. Analysis of gyrA and grlA mutations in stepwise-selected ciprofloxacin-resistant mutants ofStaphylococcus aureus., Antimicrobial Agents and Chemotherapy, 39(7), p. 1554.

Gillespie, J. J., Wattam, A. R., Cammer, S. A., Gabbard, J. L., Shukla, M. P., Dalay, O., Driscoll, T., Hix, D., Mane, S. P., Mao, C., Nordberg, E. K., Scott, M., Schulman, J. R., Snyder, E. E., Sullivan, D. E., Wang, C., Warren, A., Williams, K. P., Xue, T., Yoo, H. S., Zhang, C., Zhang, Y., Will, R., Kenyon, R. W., Sobral, B. W., 2011. PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species, Infection and Immunity, 79(11), p. 4286.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0, Systematic Biology, 59(3), p. 307.

Habib, F., Johnson, A. D., Bundschuh, R., Janies, D., 2007. Large scale genotype-phenotype correlation analysis based on phylogenetic trees, Bioinformatics (Oxford, England), 23(7), p. 785.

Hazbón, M. H., Motiwala, A. S., Cavatore, M., Brimacombe, M., Whittam, T. S., Alland, D., 2008. Convergent evolutionary analysis identifies significant mutations in drug resistance targets of mycobacterium tuberculosis, Antimicrobial Agents and Chemotherapy, 52(9), p. 3369.

Hu, M., Nandi, S., Davies, C., Nicholas, R. A., 2005. High-level chromosomally mediated tetracycline resistance in Neisseria gonorrhoeae results from a point mutation in the rpsJ gene encoding ribosomal protein S10 in combination with the mtrR and penB resistance determinants, Antimicrobial Agents and Chemotherapy, 49(10), p. 4327.

Khan, A. I., Dinh, D. M., Schneider, D., Lenski, R. E., Cooper, T. F., 2011. Negative epistasis between beneficial mutations in an evolving bacterial population, Science, 332(6034), p. 1193.

Khor, C. C., Hibberd, M. L., 2012. Host-pathogen interactions revealed by human genome-wide surveys, Trends in genetics: TIG, 28(5), p. 233.

Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., Djoumbou, Y., Eisner, R., Guo, A. C., Wishart, D. S., 2011. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs, Nucleic Acids Research, 39(Suppl 1), p. D1035.

Liu, B., Pop, M., 2009. ARDB–antibiotic resistance genes database, Nucleic Acids Research, 37(Database issue), p. D443.

Loman, N. J., Constantinidou, C., Chan, J. Z. M., Halachev, M., Sergeant, M., Penn, C. W., Robinson, E. R., Pallen, M. J., 2012. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity, Nature Reviews. Microbiology, 10(9), p. 599.

Malik, S., Willby, M., Sikes, D., Tsodikov, O. V., Posey, J. E., 2012. New insights into fluoroquinolone resistance in mycobacterium tuberculosis: Functional genetic analysis of gyrA and gyrB mutations, PLoS ONE, 7(6), p. e39754.

Manolio, T. A., 2010. Genomewide association studies and assessment of the risk of disease, New England Journal of Medicine, 363(2), p. 166.

McAleese, F. M., Foster, T. J., 2003. Analysis of mutations in the Staphylococcus aureus clfB promoter leading to increased expression, Microbiology, 149(1), p. 99.

O'Neill, A. J., Huovinen, T., Fishwick, C. W. G., Chopra, I., 2006. Molecular genetic and structural modeling studies of Staphylococcus aureus RNA polymerase and the fitness of rifampin resistance genotypes in relation to clinical prevalence, Antimicrobial Agents and Chemotherapy, 50(1), p. 298.

Patra, S. K., Jain, A., Sherwal, B. L., Khanna, A., 2010. Rapid detection of mutation in RRDR of rpo b gene for rifampicin resistance in MDR-pulmonary tuberculosis by DNA sequencing, Indian journal of clinical biochemistry: IJCB, 25(3), p. 315.

Philippe, H., Douady, C. J., 2003. Horizontal gene transfer and phylogenetics, Current Opinion in Microbiology, 6(5), p. 498.

Read, T. D., Massey, R. C., 2014. Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology, Genome Medicine, 6(11), p. 109.

Rost, B., 1999. Twilight zone of protein sequence alignments, Protein engineering, 12(2), p. 85.

Sabath, L. D., 1982. Mechanisms of resistance to beta-lactam antibiotics in strains of Staphylococcus aureus, Annals of Internal Medicine, 97(3), p. 339.

Sandgren, A., Strong, M., Muthukrishnan, P., Weiner, B. K., Church, G. M., Murray, M. B., 2009. Tuberculosis drug resistance mutation database, PLoS Med, 6(2), p. e1000002.

Schmitz, F., Sadurski, R., Kray, A., Boos, M., Geisel, R., Kohrer, K., Verhoef, J., Fluit, A. C., 2000. Prevalence of macrolide-resistance genes in Staphylococcus aureus and Enterococcus faecium isolates from 24 European University Hospitals, Journal of Antimicrobial Chemotherapy, 45(6), p. 891.

Shakil, S., Khan, R., Zarrilli, R., Khan, A. U., 2008. Aminoglycosides versus bacteria–a description of the action, resistance mechanism, and nosocomial battleground, Journal of Biomedical Science, 15(1), p. 5.

Stadler, Z. K., Thom, P., Robson, M. E., Weitzel, J. N., Kauff, N. D., Hurley, K. E., Devlin, V., Gold, B., Klein, R. J., Offit, K., 2010. Genome-wide association studies of cancer, Journal of Clinical Oncology, p. JCO.2009.25.7816.

Wattam, A. R., Abraham, D., Dalay, O., Disz, T. L., Driscoll, T., Gabbard, J. L., Gillespie, J. J., Gough, R., Hix, D., Kenyon, R., Machi, D., Mao, C., Nordberg, E. K., Olson, R., Overbeek, R., Pusch, G. D., Shukla, M., Schulman, J., Stevens, R. L., Sullivan, D. E., Vonstein, V., Warren, A., Will, R., Wilson, M. J. C., Yoo, H. S., Zhang, C., Zhang, Y., Sobral, B. W., 2014. PATRIC, the bacterial bioinformatics database and analysis resource, Nucleic Acids Research, 42(D1), p. D581.

Werckenthin, C., Schwarz, S., Roberts, M. C., 1996. Integration of pT181-like tetracycline resistance plasmids into large staphylococcal plasmids involves IS257., Antimicrobial Agents and Chemotherapy, 40(11), p. 2542.

WHO, 2010. Guidelines for ATC classification and DDD assignment.

Woźniak, M., Tiuryn, J., Wong, L., 2012. An approach to identifying drug resistance associated mutations in bacterial strains, BMC Genomics, 13(Suppl 7), p. S23.

Wu, C., Li, S., Cui, Y., 2012. Genetic association studies: An information content perspective, Current Genomics, 13(7), p. 566.