Tree-generalized hypergeometric test for detection of drug resistance-associated mutations.

Michał Woźniak*‡, Limsoon Wong†, Jerzy Tiuryn*
*Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Poland.
†School of Computing, National University of Singapore, Singapore.
‡Corresponding author: Michał Woźniak, m.wozniak@mimuw.edu.pl.

Abstract

Drug resistance in bacterial pathogens is an increasing problem, which stimulates research. In this work, in order to deepen our understanding of drug resistance mechanisms, we investigate the approach of using whole-genome sequences to identifying genetic mutations associated with drug resistance phenotypes in bacterial strains.

In particular, we present GWAMAR, the tool we have developed to support this type of analysis. As a part of this work, we also present **weighted support** (WS) and **tree-generalized hypergeometric** (TGH) score — two statistics we propose for identifying of drug resistance associations, based on phylogenetic information. 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. We present results obtained by applying GWAMAR to two datasets for *M. tuberculosis*, which demonstrate that GWAMAR can be successfully used for identification of drug resistance-associated mutations.

The software, input datasets and results are provided at the website of our project, http://bioputer.mimuw.edu.pl/gwamar.

Methodology of GWAMAR

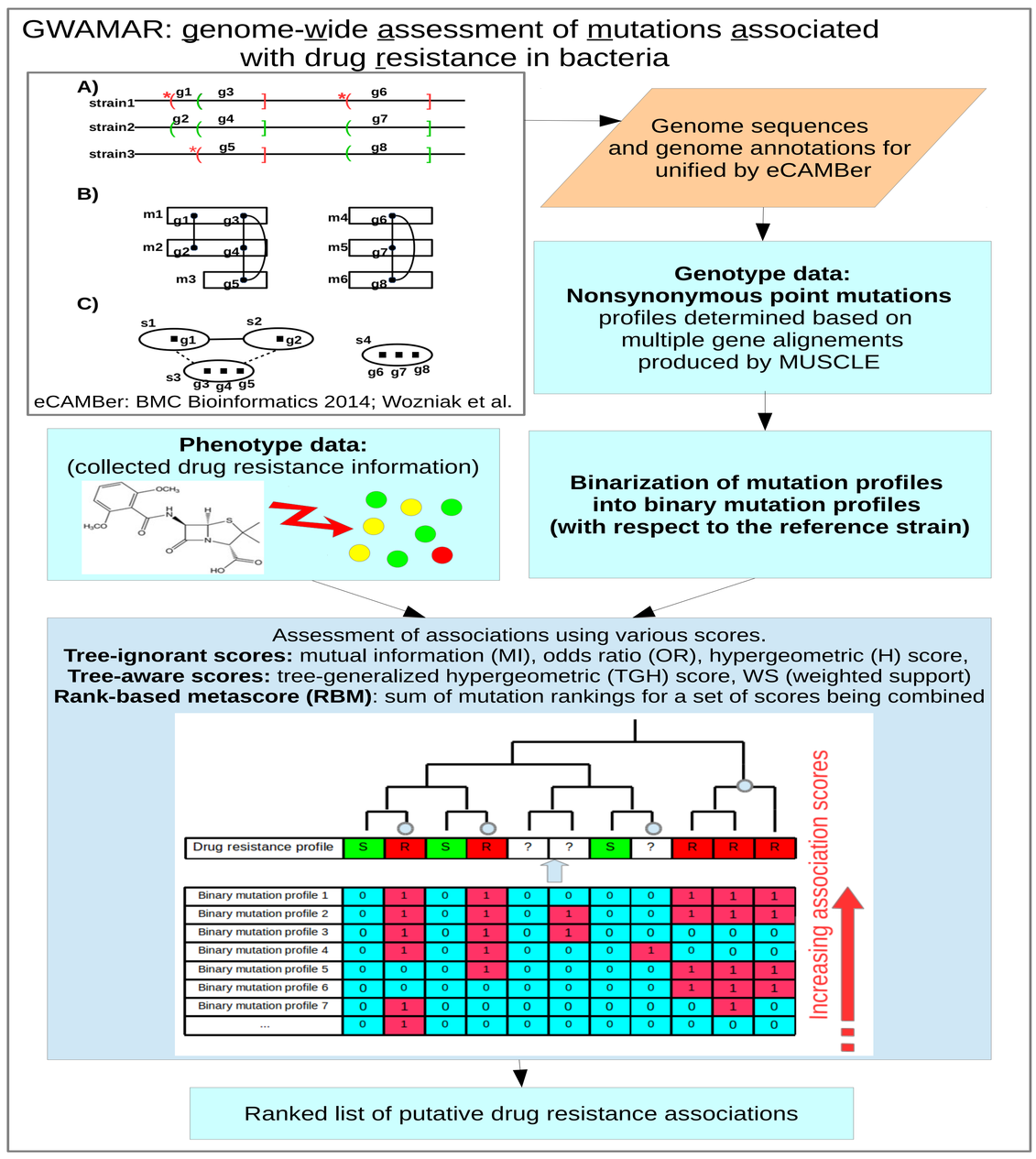
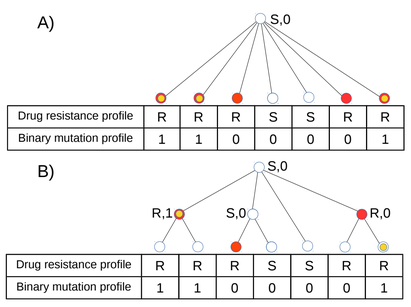


Figure 1: Schema of the pipeline of GWAMAR. For a set of considered bacterial strains, the input data for GWAMAR consists of (i) a set of mutations; (ii) a set of drug resistance profiles; and (iii) optional, phylogenetic tree for the set of bacterial strains. Typically the set of mutation profiles is 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 (MI), odds ratio (OR), hypergeometric (H) score, weighted support (WS), tree-generalized hypergeometric (TGH) score and the rank-based metascore (RBM). As a result, we obtain ordered lists of drug resistance associations, where the top-scored associations are the most likely to be real.

TGH score



A subset c of a tree nodes is a coloring, if it satisfies the following two conditions: (i) each path from a leaf to the root contains at most one node from c; (ii) each internal node in T has at least one immediate child node which does not belong to c.

(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. (B) colorings $\overline{c}$ and $\widehat{c}$ induced by the same pair of profiles but for a flat tree.

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. Then, we define the TGH score as follows:

$$\mathrm{TGH}_T(r,b) = -log\left(\frac{\sum_{i=k}^{n} B_{T,\overline{c}}(i,n)}{W_T(n)}\right). \quad (1)$$

Here, $W_T(n)$ denotes the total number of colorings of T of size n, whreas $B_{T,\overline{c}}(i,n)$ denotes the total number of colorings of T of size n, such that exactly i of their nodes are visible from coloring $\overline{c}$.

Assessment of accuracy

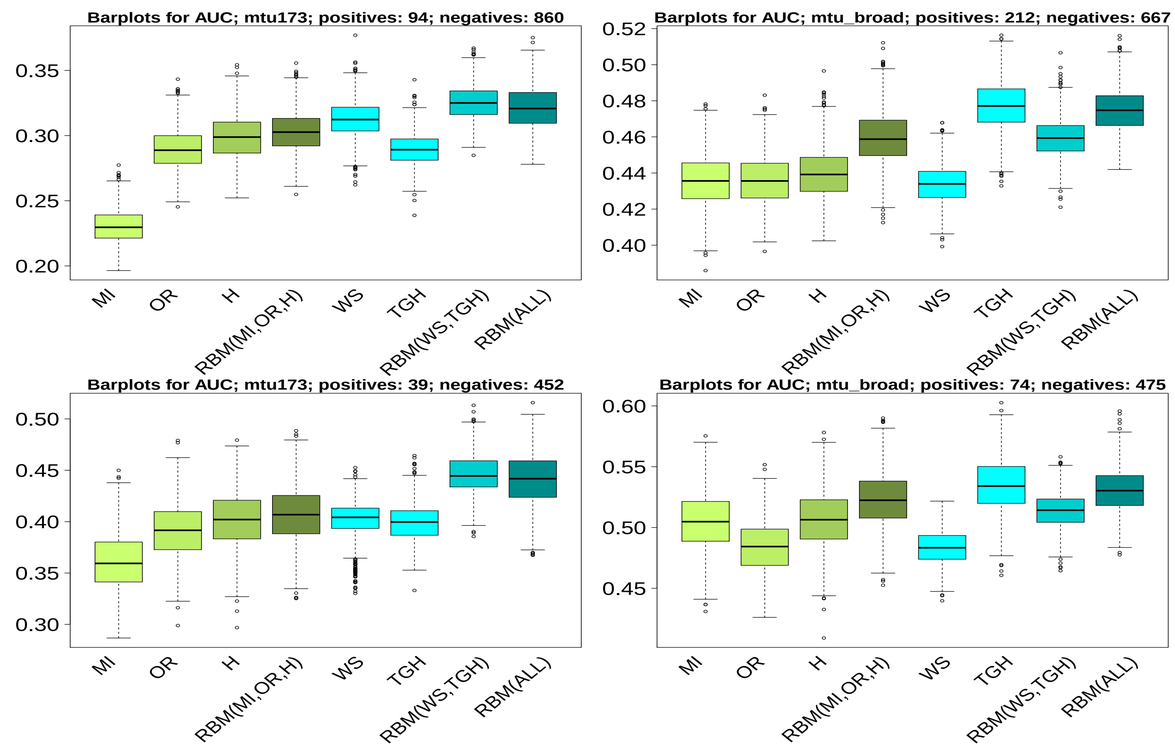


Figure 2: Tree-ignorant vs. tree aware scores: comparison of different association scores implemented in GWAMAR based on the Area Under the Curve (AUC) statistic for the precision-recall curves. Left panels present the results for the *mtu173* dataset; right for the *mtu_broad* dataset. The first row 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.

Top-scoring mutations

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

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	$H445P_4Q_2L_{27}Y_{53}R_{42}D_{25}N_7$	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

Figure 3: 25 top-scoring associations between drug resistance profiles and point mutations in the case study on 173 fully sequenced *M. tuberculosis* strains (left table) and 1398 *M. tuberculosis* strains for the Broad Institute dataset (right table). The associations are 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, and the TGH score. Lower indexes in the mutation descriptions indicate the numbers of strains possessing the corresponding amino acid or nucleotide variant.

Compensatory mutations

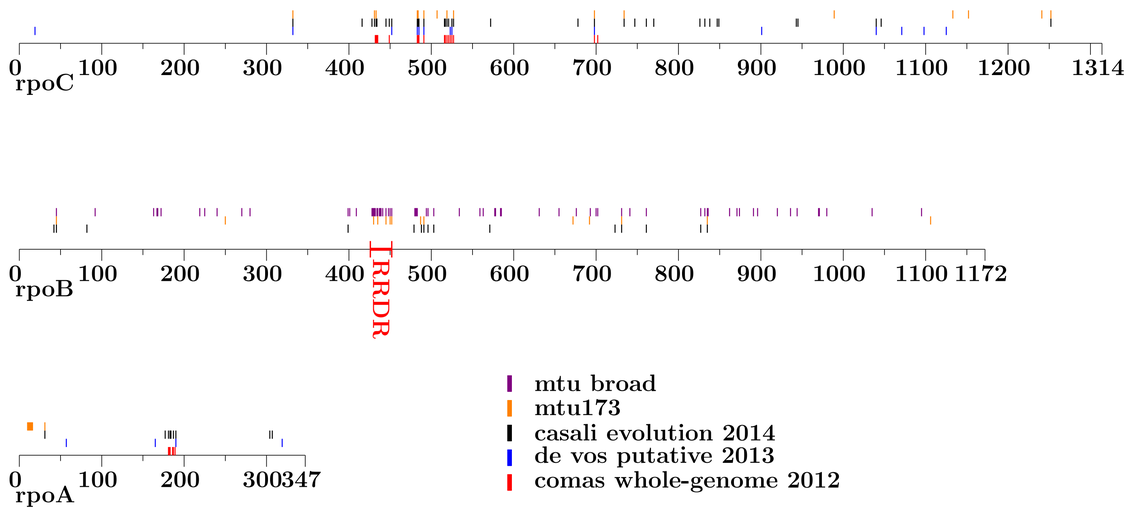


Figure 4: Comparison of the sets of putative compensatory mutations within the *rpoA*, *rpoB* and *rpoC* genes, reported in various sources and detected in our two datasets. Each mutation's position is indicated by a vertical line of the color corresponding to the source it was reported in. In particular 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.

References

[1] Woźniak M., Tiuryn J., Wong L. An approach to identifying drug resistance associated mutations in bacterial strains BMC Genomics; 2012
[2] Woźniak M., Wong L., Tiuryn J. eCAMBer: efficient support for large-scale comparative analysis of multiple bacterial strains BMC Bioinformatics; 2014
[3] Woźniak M., Tiuryn J., Wong L. GWAMAR: Genome-wide assessment of mutations associated with drug resistance in bacteria BMC Genomics; 2014
[4] Woźniak M. Computational aspects of the presence of drug resistance mechanisms PhD thesis; 2015

Acknowledgements: This work is partially supported by Polish Ministry of Science and Higher Education grants no. 2012/05/ST6/03247 and Singapore Ministry of Education Tier-2 grant no. MOE2009-T2-2-004.