Computational aspects of the presence of drug resistance mechanisms

PhD dissertation summary

Michał Woźniak

1 Introduction

Drug resistance is a broad concept used to describe situations of reduced effectiveness of a drug in curing a disease or condition. Antibiotic resistance is a form of drug resistance when some sub-population of a microorganism, typically bacteria, is capable of surviving when exposed to an antimicrobial drug. The development of drug resistance in bacteria makes antimicrobial drugs less effective and increases the cost of therapies. Since the time when first antibiotics were introduced to treat bacterial infections, due to various factors — such as irresponsible dosage of antibiotics, naturally occurring mutations, transmission of drug-resistant strains — drug resistance in pathogens has become a serious health problem. Hence, the evolution of drug resistance in bacteria is a relevant field of research in molecular biology and bioinformatics.

Different aspects of drug resistance in bacteria have been studied intensively for a few decades now (Levy and Marshall, 2004; Davies and Davies, 2010). In general, drug resistance emerges as a result of evolution which adapts bacteria to the environment with antibiotics. Some bacteria, even within wild-type communities, may have naturally increased level of drug resistance (Turnidge et al., 2006). Exposure to a drug selects these bacteria with the increased level of ability for survival in the environment with this drug. Those bacteria will produce a generation which will inherit the higher level of drug resistance. In fact, this process is so fast, that it can be reproduced and traced in a laboratory (Zhang et al., 2011) — providing a perfect example of Darwinian principles of evolution by natural selection (Sykes, 2010).
1.1 Molecular aspects of drug resistance

Antimicrobial drugs bind their molecular targets inside the bacterial cell in order to disrupt some biological processes which are essential for the bacteria (Juhas et al., 2012). In order for a drug to be effective, the following three conditions should be satisfied: (i) its drug target is in the bacterial cell, (ii) the antibiotic reaches the target in sufficient quantity, and (iii) the antibiotic is not inactivated or modified by the bacteria (Džidić et al., 2008; Blair et al., 2015).

The known drug resistance mechanisms can be categorized following the review by Wright (2011): (i) drug target modification — preventing binding of a drug; (ii) efflux — reduced accumulation of the drug inside a bacteria cell by pumping out the drug; (iii) chemical modification — modification of drug molecule by specialized enzymes; and (iv) molecular bypass — alternative metabolic pathways which can substitute for the disturbed pathways used in drug-susceptible bacteria.

On the molecular level, the process of acquisition of drug resistance is typically associated with genetic changes. These changes include chromosomal point mutations and Horizontal Gene Transfer (HGT) (Džidić et al., 2008; Davies and Davies, 2010).

Due to advances in high-throughput sequencing technologies (Loman et al., 2012; Weinstock and Peacock, 2014), the number of bacterial genome sequences available in public databases is growing rapidly. One database with bacterial genome sequences is PATRIC, developed by Gillespie et al. (2011). Notably, from June 8, 2011 to February 12, 2014, the total number of whole-genome sequences available in the PATRIC database grew from 3303 to 14114, reaching 21786 in September 24, 2014. By then, there were 1653 whole-genome sequences of E. coli and 642 whole-genome sequences of S. enterica strains available in the database (Gillespie et al., 2011).

The fast-growing number of available bacterial genome sequences enable new interesting comparative analysis of multiple bacterial strains (Binnewies et al., 2006; Hiller et al., 2007; Laing et al., 2011). In particular, it opens new opportunities to use whole-genome comparative approaches to analyse drug resistance mechanisms (Hasman et al., 2014; Alam et al., 2014).

1.2 Problems faced and approached in this work

We address in this thesis the problem of using whole-genome sequences to identify and associate genetic changes with drug-resistance phenotypes by comparative analysis of multiple closely related bacterial strains. Thus, conceptually our approach is similar to Genome-Wide Association Study (GWAS) approaches, which have been successfully applied to identify single nucleotide polymorphisms (SNPs) associated with phenotype for various human diseases (Manolio, 2010), cancer (Stadler...
et al., 2010), and intelligence (Davies et al., 2011).

We hypothesize that similar approaches, when applied to bacteria, should bring interesting results, enriching our knowledge on the molecular aspects of drug resistance. For example, better knowledge of mutations associated with drug resistance may help to design molecular test on drug resistance, such as Xpert MTB/RIF (Köser et al., 2014). This relatively cheap test allows for rapid and accurate tests for rifampicin resistance based on the presence of point mutations in *rpoB* (Boehme et al., 2010). It has become a front-line diagnostic tool in South Africa (Zumla et al., 2013). The potential of using whole-genome comparative approaches to understand bacterial drug resistance has been discussed in the recent articles by Köser et al. (2014), Hasman et al. (2014), Lázár et al. (2014), Trauner et al. (2014) and Blair et al. (2015).

We note, however, that the methodology may require some modifications to transfer it to bacteria. For example, horizontal gene transfer (HGT) plays an important role in the development of drug resistance in bacteria (Warnes et al., 2012); thus it may be needed to focus not only at SNPs, but also at gene gain/losses.

One challenge we faced during this project was caused by inconsistent and poor-quality annotations of bacterial strains available in public databases. It has been argued in various articles that these inconsistencies are due to different annotation methodologies used by different sequencing laboratories (Overbeek et al., 2007; Dunbar et al., 2011; Chai et al., 2014). It has also been shown, that poor-quality annotations may complicate or bias the comparative analysis of bacterial strains (A Palleja et al., 2008; Cock and Whitworth, 2010; Dunbar et al., 2011; Yu et al., 2011; Wood et al., 2012). The tools we developed to tackle this problem are presented in the second chapter of the thesis.

The next challenge we faced was to collect sufficient amount of data on outcomes of drug susceptibility tests for different strains. These outcomes would constitute the phenotype data for further analysis. It turned out that this data is spread throughout the literature and databases, thus not easy to gather. We made the collected information publicly available at the website of our project, http://bioputer.mimuw.edu.pl/gwamar.

Finally, having collected the phenotype and genotype, we approach the problem of associating the identified genetic variations with the drug-resistance phenotypes. In order to achieve this, we implemented several association scores, including two new association scores, called weighted support (WS) and tree-generalized hypergeometric (TGH) score. We also propose a new association score, called Rank-based metascore (RBM) for combining multiple scores into one in order to compromise between different approaches used to define different scores.
The concept of genome annotation may refer to many different aspects of attaching biological information to genome sequences, such as: identifying gene locations (Karp et al., 2007), assigning functions to genes (Richardson and Watson, 2013), and assigning network context to gene products (Kasif and Steffen, 2010).

In this work, we focus on identifying locations of protein-coding genes. The translation unit of a protein-coding gene is a continuous fragment of DNA, of length divisible by 3. It begins with a start codon (typically ATG, but also GTG and TTG are common) at the translation initiation site (TIS), and ends with a stop codon (TAA, TAG or TGA). An open-reading frame (ORF) is any fragment of DNA which satisfies the above conditions, thus it has the potential to code for a protein. However, the presence of an ORF does not imply that the region is translated.

We use the term gene annotation (or ORF annotation) to refer to genome coordinates of the translation unit of a protein-coding gene from its TIS (alternatively called gene start) to the nearest stop codon (alternatively called gene end). Note that each ORF annotation is unambiguously determined by specifying strand and position of its start codon.

Due to gene annotation inconsistencies, multiple ORF annotations — corresponding to the same gene in different strains — may suggest different lengths of the gene. In order to account for this situation, we generalize the concept of a gene annotation. We introduce a term we called a multigene annotation to represent multiple gene annotations sharing the same stop codon. Obviously gene annotations can be viewed as multigene annotations with just one element.

In this work we present CAMBer and its highly optimized version, eCAMBer. These are tools that we have developed to improve the consistency and overall quality of bacterial genome annotations by comparative genome annotation.

In its key step, called the closure procedure, eCAMBer tries to transfer gene annotations among all considered bacterial strains. The underlying idea behind the efficient implementation of the procedure in eCAMBer is to avoid redundant BLAST queries. Moreover, eCAMBer supports multiple-threading for all its procedures. This allows eCAMBer to be much faster than CAMBer and its other competitors — Mugsy-Annotator (Angiuoli et al., 2011) and the GMV pipeline Wall et al. (2011) — making it applicable to datasets comprising hundreds of bacterial strains.

For example, on the dataset of 41 strains of E. coli, computations took less than two hours (using only one processing thread), whereas Mugsy-Annotator (the fastest competitor) took more than 19 hours.

Furthermore, eCAMBer tries to resolve annotation inconsistencies in order to produce more accurate annotations. For this purpose, it implements a majority
voting-like approach for selecting the most reliable TISs and implements a procedure for identification and removal of gene families which are likely to be propagated annotation errors.

In order to assess the impact of using eCAMBer on the quality of annotations, we applied it on the dataset of 20 *E. coli* strains, comprising genome sequences and annotations from the PATRIC database. As a gold standard for this dataset we used genome annotations manually curated by biologists. The results showed that eCAMBer improved the quality of the input annotations.

Results presented in the thesis also show, that eCAMBer outperforms its competitors, Mugsy-Annotator and the GMV pipeline, in terms of improving quality of annotations. In particular, when run on genome annotations generated by Prodigal ([Hyatt et al., 2010](#)), for the set of 20 *E. coli* strains, eCAMBer increased the f1 statistic of initial annotations from 0.764 to 0.775, whereas the application of GMV improved it only by 0.001 and the application of Mugsy-Annotator even decreased it.

Although the development of eCAMBer was motivated by our research on identifying genetic variations associated with drug resistance, it may be used in other contexts where high-quality annotations of bacterial strains are needed. We expect eCAMBer to be a valuable tool for the research community working on comparative analysis of multiple bacterial strains.

We expect that, with the increasing amount of genomic data being generated, the need for similar tools will continue to grow. For example, to date, there are already about three thousand sequenced *S. aureus* genomes. Comparative analysis of such large datasets may be difficult for all currently available tools. One promising approach to achieve higher efficiency may rely on the idea of compressive genomics, which can greatly speed up running time of BLAST queries when the target genome sequences are highly similar, as in the case of closely related bacterial strains. We leave this, however, as one of the potential directions for further research.

The tools, case-study input data and the obtained results are available at the website of this project, [http://bioputer.mimuw.edu.pl/ecamber](#).

### 3 Results for drug resistance-associated mutations

In this work, we present GWAMAR, the tool we have developed for genome-wide assessment of mutations associated with drug resistance.

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.

The input genomic data comprises genome sequences and annotations for a set of bacterial strains. The input phenotype data consists of information on drug susceptibility collected for the set of strains in question.

After the preprocessing step, the genotype data consists of a set of genetic mutations of three types:

- point mutations,
- 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:

\[ v : S \rightarrow \Sigma. \tag{1} \]

For a given mutation profile \( v \), its corresponding binary mutation profile

\[ b_v : S \rightarrow \{0, 1, ?\}, \tag{2} \]

is defined as follows:

\[ b_v(S) = \begin{cases} 
? & \text{if } v(S) = ? \\
0 & \text{if } v(S) = v(S^*) \\
1 & \text{otherwise} 
\end{cases} \tag{3} \]

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:

\[ r : S \rightarrow \{S, I, R, ?\}. \tag{4} \]

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.

The essential step of GWAMAR is scoring of the set of mutations with regard to the phenotype data. The goal of this procedure is to achieve a sorted list of
associations, such that, the associations with the highest score are most likely to be real.

In this order, GWAMAR implements several statistics, such as mutual information, hypergeometric score, and odds ratio. As a part of this work we investigate the potential of scoring which employs phylogenetic information, represented as a tree, to strengthen the signal in data, which comes from the presence of the underlying association. We propose two such statistics, weighted support (WS) and the tree-generalized hypergeometric (TGH) score. Since it is not clear that trees are the best data structures to represent the phylogenetic information, one potential direction of the research could be to test the usability of other data structures, such as networks, to represent the phylogenetic information.

In order to test our approach, we prepared one dataset for \textit{S. aureus} and two datasets for \textit{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, we identify some putative associations, which haven’t been yet reported. These \textit{in silico} predictions may attract the attention of the experimental research community to test them experimentally in the labs.

Even though, in this work we focus solely on the problem of identifying drug resistance-associated mutations in bacteria, very similar approaches can be developed to study drug resistance in viruses, parasites or cancer. Moreover, in our opinion, the tool could be successfully applied to other types of phenotype data, such as virulence.

Finally, we expect that, with the increasing amount of genomic data being generated, more studies on whole-genome comparative approaches will be published in the context of drug resistance.

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

4 Authors and articles

Most of the results presented in the thesis have been published in peer-reviewed articles.

In the first chapter, we briefly review the current background knowledge of the drug resistance mechanisms. A broader coverage of the biological background can be found in the review articles by Levy and Marshall (2004), Đžidić et al. (2008),
Davies and Davies (2010) and by Wright (2011).

In the second chapter, we present our work on supporting comparative analysis of multiple bacterial strains. We present and describe the methods and the software we have developed to support that analysis. The presentation is based on our three articles describing CAMBer (Woźniak et al., 2011a), CAMBerVis (Woźniak et al., 2011b), and eCAMBer (Woźniak et al., 2014b).

In the third chapter, we present our work on detection of drug resistance-associated mutations based on comparative analysis of whole-genome sequences of multiple bacterial strains. In particular, we present GWAMAR, the tool we have developed to support that analysis. The presentation is based on our two articles (Woźniak et al., 2012) and (Woźniak et al., 2014a).
REFERENCES


