Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Microbiology

Bioinformatics of antimicrobial resistance in the age of molecular epidemiology

Andrew G McArthur and Gerard D Wright

Antimicrobial resistance is a global health challenge and has an evolutionary trajectory ranging from proto-resistance in the environment to untreatable clinical pathogens. Resistance is not static, as pathogenic strains can move among patient populations and individual resistance genes can move among pathogens. Effective treatment of resistant infections, antimicrobial stewardship, and new drug discovery increasingly rely upon genotype information, powered by decreasing costs of DNA sequencing. These new approaches will require advances in microbial informatics, particularly in development of reference databases of molecular determinants such as our Comprehensive Antibiotic Resistance Database and clinical metadata, new algorithms for prediction of resistome and resistance phenotype from genotype, and new protocols for global collection and sharing of high-throughput molecular epidemiology data.

Address
M.G. DeGroote Institute for Infectious Disease Research, Department of Biochemistry and Biomedical Sciences, DeGroote School of Medicine, McMaster University, 1280 Main Street West, Hamilton, ON, Canada L8S 4K1

Corresponding author: McArthur, Andrew G (mcarthua@mcmaster.ca)

Current Opinion in Microbiology 2015, **27**:45–50

This review comes from a themed issue on **Antimicrobials**

Edited by **Paul M Dunman** and **Andrew P Tomaras**

http://dx.doi.org/10.1016/j.mib.2015.07.004

Introduction

Antimicrobial resistance (AMR) is among the most pressing health crises of the 21st Century [1,2•]. The low estimate for global deaths due to drug resistant microbial infections now is 700,000 annually, but without intervention is predicted to be 10 million annually by 2050, with an associated global cost of 100 trillion USD and global 2–3.5% drop in gross domestic product (GDP) (UK Review on Antimicrobial Resistance, amr-review.org). These numbers are staggering, yet antibiotics are important not only to treat bacterial infection, but also enable much of modern medicine. Interventions such as heart surgery, joint replacements, cancer chemotherapy, and transplantation all require robust prophylactic control of infection and thus require antibiotics. In short, without antibiotics we lose much of modern medicine, and increasingly we have fewer antibiotics because of resistance. We are at risk of entering a 'post-antibiotic era' [1].

Despite the importance of resistance to health, the field has been slow to take advantage of genome scale tools. Rather, laboratory-based phenotype criteria dominate the epidemiology of antibiotic action and effectiveness. As a result, there is a poor understanding of which antibiotic resistance genes are in circulation, which ones are a threat, and how clinicians and public health workers can manage the crisis of resistance. Importantly, there is generally no effort to link phenotype to actual resistance genes. In the absence of a robust pipeline of new drugs coming to market, understanding the genomic basis of resistance and its movement through bacterial and patient communities is essential for judicious management of increasingly scarce antibiotics and to guide new drug discovery. Fortunately, DNA sequencing is rapidly decreasing in cost and as such we are on the cusp of an era of high-throughput molecular epidemiology. Consequently, we need tools for rapid, accurate analysis of DNA sequence data for the genetic underpinnings of antibiotic resistance. Additionally, we need a long-term plan for best use of these data when DNA sequencing of pathogens becomes commonplace at every hospital, clinic, and outbreak. In an effort to address this problem, we have created the Comprehensive Antibiotic Resistance Database (CARD; arpcard.mcmaster.ca) [3•].

The CARD is a rigorously curated collection of known resistance determinants and associated antibiotics, organized by the Antibiotic Resistance Ontology (ARO), a theoretical framework for organizing antibiotic resistance information. The CARD also includes a predictive tool, the Resistance Gene Identifier (RGI), which predicts known antibiotic resistance elements from genome sequence data. These first steps in the development of the CARD are already proving highly useful to a number of researchers and public health investigators, yet are in many ways reflective of traditional design of biological databases. Our ultimate goal is to transform the CARD from a traditional biological database to a powerful tool for wide-spread generation, analysis, and sharing of data on the molecular epidemiology of antibiotic resistance.

High quality reference data

The use of genome sequencing to aid diagnostics and affect clinical outcomes involves the methods of comparative

genomics and requires high quality reference data. There have been attempts to build databases of antibiotic resistance elements in the past. The Bush–Jacoby β-lactamase list is exclusively focused on compiling an inventory of a small subset of resistance genes [4], the ARDB was an effort to generate a preliminary antibiotic resistance gene ontology and gene catalog, but has not been updated since 2009 [5], and BacMet catalogs metal and biocide resistance genes that may additionally drive antibiotic resistance [6]. All of these are laborious manual cataloging efforts, making their translation to high-throughput diagnostics difficult. The first stages of the CARD have not been dissimilar. To date, the data in CARD are hand-curated to ensure high quality entries and includes genes and their products, plasmids, compounds (antibiotics and other relevant chemicals such as inhibitors of resistance), classification of pathogens, and associated literature references. The resistance gene information in CARD is linked to entries in PubChem, PubMed, GenBank and the PDB (Protein Data Bank), providing a seamless association with literature and chemical and drug databases. Key to this development of the CARD has been the establishment of the Antibiotic Resistance Ontology (ARO). Ontologies are controlled vocabularies that form the foundation of genome bioinformatics; they provide a robust vocabulary for genes and their products that link them to their activities and enable robust investigation of molecular data [7]. The ARO thus provides a unifying language specific to antibiotic resistance that enables codification of resistance and target genes, drugs, and molecular activities germane to the field. It is a critical step toward development of standards for data sharing among disparate research teams and databases working on AMR and can be readily downloaded from the CARD website.

The CARD seeks to acquire curated information on resistance mechanisms, genes, and their targets to create a rich resource for development of algorithms for prediction of antibiotic resistance from genome sequence. While curation of molecular sequence data (i.e. resistance genes) is manageable, complex metadata on prevalence and drug resistance profiles (e.g. minimum inhibitory concentrations, association with bacterial genera and species) is often buried in the scientific literature. Yet, from a predictive and clinical perspective it is critical that we know the distribution of resistance genes around the globe, among pathogens, among infection types, environmentally, and in relation to patient demographics, as well as the range of antibiotics they act upon within each pathogen. Thus, the curation of knowledge and metadata from the published literature is a key challenge for AMR bioinformatics. However, development of algorithms for extraction of association and interaction information from the literature is a major area of research in the field of BioCuration and is actively used to enrich a number of biological databases [8–12]. In particular, algorithms for extraction of gene–gene or gene–compound interactions are making important contributions to genome annotation [13], construction of gene regulatory networks [9], and cataloging of drug-drug interactions [11] and we are pursuing similar algorithms to extract antibiotic resistance association and interaction data from the scientific literature for the CARD. For example, while manual curation of the CARD has been able to ensure inclusion of nearly all resistance genes, data on the range of antibiotics affected by these genes has proven difficult to compile given the volume of literature. While the CARD includes this information at a broad level, such as knowledge that aminoglycoside acetyltransferases (AACs) confer resistance to aminoglycoside antibiotics, it has not proven feasible for manual curation to compile the degree of resistance conferred by individual AACs to individual aminoglycosides. This level of detail, important for prediction of the antibiogram (the phenotypic resistance profile of a given strain) from genotype and varying among pathogens, does not currently exist in any AMR database. Similarly, text mining is needed to compile association of AMR genes among pathogens and regions, in relation to epidemiology, clinical metadata, and even means of dissemination (e.g. plasmid, horizontal gene transfer, transposable elements). Text mining will thus enable construction of a matrix of gene-gene associations, both within genomes or plasmids but also among pathogens, allowing analysis of genome sequences within a probabilistic framework, where algorithms can use association matrixes to predict antibiogram in the context of missing data based on detection of strongly associated molecular determinants.

Prediction of resistome and antibiogram

The increasing use of next generation whole genome sequencing (WGS) and whole community sequencing (WCS; a.k.a. metagenomics) is revolutionizing medical microbiology, resulting in a paradigm shift from phenotype to genotype-based diagnostics of resistance [14,15•]. We now have the capacity to rapidly and affordably sequence whole genomes to unlock the biology and epidemiology of pathogens. Equally important, we are on the cusp of an age where these technologies and research methods can translate to improved clinical outcomes, with rapid detection of drug resistance genes in infectious pathogens. The challenge is no longer acquisition of data, but analysis of data and we are facing bottlenecks in both the development of informatics tools to assess this flood of genomic information and in the training of the next generation of biomedical researchers and clinicians capable of harnessing these high-throughput data. Within the CARD, the Resistance Gene Identifier (RGI) currently provides prediction of resistance genes from DNA sequences based upon the curated data available in the CARD. The RGI analyzes genome sequences or assemblies relative to the CARD using BLAST and provides a detailed output of predicted antibiotic resistance genes and targeted drug classes. This includes