Review

Designing a low-cost drug resistance database for viral hepatitis

Carla Kuiken1*

1HIV database project, Los Alamos National Laboratory, Los Alamos, NM, USA

*Corresponding author e-mail: Kuiken@lanl.gov

Introduction

Treatment of patients infected with hepatitis B with potent antiviral drugs is well established and a large array of resistance-associated mutations has been documented for HBV (see for example [1] for a review). For hepatitis C infection, treatment with these drugs has started only recently, but it is already clear that resistance will also be a major issue in HCV treatment [2,3]. For successful treatment, it is vital to know patients’ resistance profile for several reasons. First, when a drug or regimen loses effectiveness, the cause is often the development of resistance; however, the next treatment step often depends on the mutation or combination of mutations that are present in the viral quasispecies [4]. Knowledge of how to adjust treatment regimens when certain resistance patterns develop is usually available in-house in big hospitals, where specialists are available for each virus, but not as easy to come by in smaller and under-resourced clinics, which often serve poorer and sicker patients with higher comorbidity. Second, the field of drug resistance evolves fast that even clinicians in well-resourced hospitals can find it difficult to keep abreast of recent developments. Even clinicians who are well-versed in hepatitis B drug resistance issues might not have equal mastery of other viruses that often infect the same patients. The increased mobility of populations from under-resourced countries, where treatment lags behind fast-evolving standards of care, mixes viruses with different treatment histories, further complicating the resistance profiles. Treating patients in developing countries is an even bigger challenge, because the frequency patients can be tested for resistance, as well as the availability of possible alternate drugs, might be limited.

In the absence of a central and up-to-date repository for drug resistance information, clinicians miss an important tool in their arsenal that could greatly advance treatment and avoid medication errors that can easily cause further resistance. In addition, it would be very helpful for drug resistance researchers to have quick, easy access to such a repository. For HIV type-1, there is a very well-maintained and heavily used resistance database [5]. The group maintaining this database has also developed a large number of tools to analyse resistance patterns and outline possible further treatment regimens for a given resistance pattern [6,7]. Setting up and maintaining a resistance database is a major project. The Stanford HIV-1 Drug Resistance Database has existed for almost 10 years, so it has a stable infrastructure. It employs five people who annotate and maintain the database and develop new tools to analyse the data. However, although a specialized, stably funded resistance database for each significant pathogen is definitely optimal, for most virus research fields this is not currently feasible.
Existing infrastructure

In this paper, we outline a relatively cheap and simple way to create a maintainable antiviral drug resistance mutation database for less well-funded viruses. The proposed database will be built on top of existing sequence databases [8] and relies, in part, on community annotation. Once the database infrastructure is in place, we suspect it will be feasible to create and communally maintain an up-to-date viral resistance database. The target audiences for the database are in small communities of highly motivated virus researchers and clinicians, who will receive immediate benefit from the database as well as appreciation and acknowledgement for their contributions. Initially, we will focus our efforts on the HBV and HCV communities, where the need for a way to track resistance mutations is particularly acute [9,10]. The Los Alamos HCV database is currently unfunded, but it is still being maintained and work is ongoing to further automate maintenance so the database becomes increasingly maintenance-free.

The HIV and HCV databases in Los Alamos can be adapted to store information on other viruses and the associated tools can then be used to analyse data from these viruses with minor adaptations. The Los Alamos project has received funding to implement such an adaptation for the haemorrhagic fever viruses, which include Arenavir-, Bunyavir-, Filovir-, some Flavivir- and Togaviruses. This new database is almost organism-independent and functions by using the annotation of International Nucleotide Sequence Database Collaboration (INSDC) genome reference sequences (this database contains a single, carefully annotated representative sequence for each species [11]) to direct the analyses for all viruses in that species. To know which position in each sequence corresponds to which mutation, the sequences must be aligned unambiguously and their positions must be known. For this, it seems the method followed by the HCV and HIV databases is the easiest to implement. This method is based on a model sequence, an internal database representation that is able to accommodate all sequences that are stored. All sequences, including the reference sequence, are internally aligned against this model sequence and thus the position of all nucleotides is known [12].

Creating a resistance database

A resistance database could be added to this infrastructure as follows. The resistance information in the database would be modelled on existing resistance databases, although it will probably be simplified for ease of maintenance. Data stored will include the compound name and synonyms, phenotypic resistance information when available, the resistance-associated mutations (location and ‘wild-type’ versus ‘mutant’ amino acid and codon), interactions with other mutations, evidence (possibly a text field), references, and annotator. The amount of data that needs to be annotated is quite small, on the order of about 10 pieces of information for each new resistance mutation. With this information, the frequencies of mutations in the sequence database can be retrieved and displayed, along with any available associated data.

Access to the database will be open, but annotation will not be anonymous. We expect the number of annotators for each virus to be small (approximately 5–10 people). There will also need to be one or at most two editors for each virus, who will look at new entries, judge their quality and will add or remove annotators when required. The National Center for Biotechnology Information ‘RefSeq’ database works on this basis, with one editor who recruits unpaid annotators who are specialists on one area. The database itself will be centrally located and maintained. Our experience indicates that once a database is properly set up and the teething problems have been dealt with, it does not require much attention other than annotation.

The sequence data and metadata originate from two sources: genomic database downloads (mostly coming from the INSDC, which Genbank, European Molecular Biology Laboratory and DNA Data Bank of Japan share [13]), and manual annotation. The procedure for INSDC downloads of nucleotide sequences has been operational for many years at Los Alamos and is mostly automated. It can be easily generalized to represent other viruses. Work is currently in progress to allow modifications to the database schema to translate with minimal effort to the corresponding modifications in the search and update interfaces.

Relevant metadata for the sequences can include epidemiology data, demographic data, health status and coinfection information, and therapy (treatment) data. Epidemiology data, demographic data, health status and coinfection information are currently being annotated for the American HCV and HIV databases. Treatment data are currently not annotated, although there is a model for storing these, based on that used by the HIV drug resistance database at Stanford University. In collaboration with the Los Alamos Institutional Review Board, the degree of detail for these data has been limited to preclude possible identification of individual patients.

Annotations the database

Community annotation of technical subjects is a relatively new development. The front runner in this field is of course Wikipedia. Annotation projects can be characterized in many different ways, but one important dimension is the expert level of their annotators. There is an ordering along this axis in a series of community
annotation projects that range from Wikipedia (anyone can annotate), via Citizendium (annotators are identified by name), Scholarpedia (annotators are invited or elected), to the PLoS and other open-access journals (submissions are peer reviewed). In physics, the ‘ArXiv’ [14] has been a central repository for preprints and some influential papers are never published anywhere else; submissions are only screened for general suitability and sensibility (for example, creationist papers are removed), but otherwise ArXiv relies on its readers to critique the manuscripts.

Community annotation will form a vital part of the new database because continuous funding for annotation is very difficult to secure. Clinicians and researchers interested in drug resistance mutations for hepatitis B and C have a strong interest in organizing and centralizing the information that will be covered in this database. Furthermore, annotating the resistance information in the database will involve a small amount of work, unlike annotating the function of a gene or protein, which can take weeks or even months.

However, any community annotation project needs one or several editors to keep track of the quality and speed of annotation, and to add or remove annotators when necessary. We would anticipate that these editors, whose time investment will likely be on the order of an hour a week or so, can be invited to jointly publish the initial announcement of the database once it has been created and thus benefit from increased use and referencing of the database. Annotators will be asked to sign up using their real name and requests for accounts will be monitored. The account can also be closed if annotators are found to generate too many mistakes; this will most likely only happen after complaints from editors or users. Annotators’ names will be visible when accessing their information along with the number of edits they made. Both editing and annotating will likely come to be seen as activities that serve the community and the clinical care of patients, and therefore as legitimate enhancements of a resume.

Good examples exist of how this resistance database could be searched. The resistance mutation database will clearly need a new interface that is separate from the existing sequence search interfaces. It will allow searches on the basis of drug, drug type, mutation type, location, interactions, coinfection and genotype, among others. Available metadata, for example, geographical, temporal and clinical, can be displayed and used to limit the results. Easy-to-read graphical displays will be designed or copied from existing resources, for example, the Stanford HIV database and a new tuberculosis drug resistance database maintained at Johns Hopkins University, Baltimore, MD, USA [13].

Conclusion

The creation of a publicly accessible site for centralized storage, analysis and tracking of antiviral drug resistance mutations would improve treatment of viral infections. It would allow clinicians to avoid ineffective treatment courses and therefore help prevent further resistance development. It would provide patients with more effective and less protracted and toxic treatment regimens, and reduce the use of unnecessary and very costly drugs. It would be particularly valuable for viral treatment in disadvantaged and under-resourced settings, and will be beneficial for researchers, clinicians and patients.

Acknowledgements

We thank our colleagues at the HIV databases for their help and support. Funding for this work was provided by the Division of AIDS of the National Institutes of Health, Bethesda, MD, USA (HIV database project, contract number Y1AI830902).

Disclosure statement

The author declares no competing interests.

References


Accepted for publication 18 September 2009