The Journal of Open Source Software

sierra-local: A lightweight standalone application for drug resistance prediction

Jasper C. Ho[1], **Garway T. Ng**[1], **Mathias Renaud**[1], **and Art F. Y. Poon**[1, 2, 3]

1 Department of Pathology and Laboratory Medicine, Western University, London, ON, Canada **2** Department of Microbiology and Immunology, Western University, London, ON, Canada **3** Department of Applied Mathematics, Western University, London, ON, Canada

DOI: 10.21105/joss.01186

Software

- Review
- Repository
- Archive

Submitted: 26 November 2018
Published: 25 January 2019

Summary

Genotypic resistance interpretation systems for the prediction and interpretation of HIV-1 antiretroviral resistance are an important part of the clinical management of HIV-1 infection. Current interpretation systems are generally hosted on remote webservers that enable clinical laboratories to generate resistance predictions easily and quickly from patient HIV-1 sequences encoding the primary targets of modern antiretroviral therapy. However they also potentially compromise a health provider's ethical, professional, and legal obligations to data security, patient information confidentiality, and data provenance. Furthermore, reliance on web-based algorithms makes the clinical management of HIV-1 dependent on a network connection. Here, we describe the development and validation of *sierra-local*, an open-source implementation of the Stanford HIVdb genotypic resistance interpretation system for local execution, which aims to resolve the ethical, legal, and infrastructure issues associated with remote computing. This package reproduces the HIV-1 resistance scoring by the web-based Stanford HIVdb algorithm with a high degree of concordance (99.997%) and a higher level of performance than current methods of accessing HIVdb programmatically.

Background and Rationale

Genotype-based prediction of human immunodeficiency virus type 1 (HIV-1) drug resistance is an important component for the routine clinical management of HIV-1 infection (Günthard et al., 2014; Tural et al., 2002). Detecting the presence of viruses carrying mutations that confer drug resistance enables physicians to select an optimal drug combination for that patient's treatment regimen. Furthermore, genotyping by bulk sequencing is a cost-effective alternative to the direct measurement of drug resistance from culturing virus isolates in a laboratory (Mayer, Hanna, & D'Aquila, 2001). Provided access to affordable bulk sequencing at an accredited laboratory for clinical microbiology, the interpretation of HIV-1 sequence variation is the primary obstacle to utilizing resistance genotyping for HIV-1 care.

Fortunately, there are several HIV-1 drug resistance interpretation algorithms that can be accessed at no cost through web applications or services hosted by remote network servers, such as the Standard University HIV Drug Resistance Database (HIVdb) (T. F. Liu & Shafer, 2006), Agence Nationale de Recherche sur le SIDA (ANRS) AC11 (Meynard et al., 2002), and Rega Institute (Van Laethem et al., 2002) algorithms. The Stanford HIVdb

interpretation system can be accessed either through a web browser at http://hivdb.stanford.edu/hivdb or programmatically through its Sierra Web Service (Tang, Liu, & Shafer, 2012), which requires the transmission of an HIV-1 sequence from a local computer over the network to the remote server. This is a convenient arrangement for clinical laboratories because there is no need to install any specialized software, web browsers are ubiquitous and most users are familiar with submitting web forms. Alternatively, a laboratory may host an instance of the HIVdb Sierra Web Service itself, which was recently made possible with the release of the Sierra source code. This approach, however, requires the configuration of a web server, the Apache Tomcat web container, and a large number of Java libraries.

There are a number of disadvantages to accessing interpretation systems over a network connection. First, HIV-1 sequences are sensitive patient information, not only because infection with HIV-1 remains a highly stigmatized condition, but also because sequence data have been used as evidence in the criminal prosecution of individuals for engaging in sexual intercourse without disclosing their infection status, leading to virus transmission (Bernard, Azad, Vandamme, Weait, & Geretti, 2007). Once sequence data have been transmitted to a remote server, one cedes all control over data security. Preventing the onward distribution of the data and deleting the data once the analysis is complete, for instance, is entirely the responsibility of the system administrators of the host server. Furthermore, unless the host server employs a secure transfer protocol, the unencrypted data are transmitted in the clear between a number of intermediary web servers, exposing these data to a 'man-in-the-middle' attack (Patil & Seshadri, 2014).

Second, the algorithm hosted on the server is effectively a black box — one has no insight into how resistance predictions are generated. Even if a version of the algorithm has been released into the public domain, one cannot be certain that the exact same algorithm was applied to their transmitted data. Importantly, different versions of a given algorithm can output significantly different resistance predictions, with the general trend being an increase in both resistance scores and predicted resistance levels (Hart, Vardhanabhuti, Strobino, & Harrison, 2018). In addition to contributing to inconsistencies in algorithm outputs, this makes it difficult to track data provenance, *i.e.*, the historical record of data processing, that has become recognized as a critical gap in the workflows of clinical laboratories. For instance, the College of American Pathologists recently issued new accreditation requirements stipulating that clinical laboratories must track the specific version of software programs used to process patient data (Aziz et al., 2015). Thus, a reliance on web-based systems creates significant issues for the reproducibility and quality assurance of clinical workflows. The Stanford HIVdb web service (Sierra (Tang et al., 2012)), for instance, automatically utilizes the most recent version of the HIVdb algorithm. While this constraint ensures that users employ the most up-to-date algorithm, it also introduces hidden changes to clinical pipelines, which may have been locally validated on older versions of the algorithm.

Third, dependence on a web resource may cause problems when the laboratory cannot access the host server, either due to local or regional network outages, or because the host server is malfunctioning or offline. In our experience, the web servers hosting the more popular HIV drug resistance interpretation algorithms such as the Stanford HIVdb database are reliable and well-maintained. However, it is not unusual for other web-based algorithms to be relocated or go offline when the developers move to other institutions or lack the resources to maintain the service.

One of the important features of the Stanford HIVdb algorithm is that it is regularly updated and released into the public domain in a standardized XML-based interchange format — the Algorithm Specification Interface version 2 (ASI2) format (Betts & Shafer, 2003) — that was formulated and published by the same developers in conjunction with the Frontier Science Foundation. Here, we describe the implementation and validation of *sierra-local*, an open-source Python package for local execution of the HIVdb algorithm

in the ASI2 format. This package utilizes, but does not require, a network connection to synchronize its local ASI2 file and reference data with the latest releases on the Stanford HIVdb web server. Our objective was to release a lightweight alternative to transmitting HIV-1 sequences to the HIVdb web server that minimizes the number of software dependencies, and that produces the exact same interpretations as the Sierra web service for all available HIV-1 sequences in the Stanford database.

Validation

We obtained the entirety of the genotype-treatment correlation datasets available through the Stanford HIV Drug Resistance Database (HIVdb (Shafer, 2006)) on May 7 2018. After screening for invalid data, the resulting dataset contained 103,711 HIV-1 protease, 110,222 reverse transcriptase and 11,769 integrase entries. We scored these data with both *sierra-local* and SierraPy (version 0.2.1, https://github.com/hivdb/sierra-client) using the HIVdb version 8.5 algorithm on both platforms. Because the algorithm was updated to version 8.6.1 during the validation experiments, we used the newer version for the HIV-1 integrase data sets since the update mostly affected the interpretation of mutations within this region. Out of the total 226,702 sequence records, the predicted resistance scores were completely identical in 226,696 (99.997%).

In addition, we retrieved 7 population-based HIV-1 *pol* datasets from Genbank using the NCBI PopSet interface (http://www.ncbi.nlm.nih.gov/popset). These datasets were selected from the most recent uploads of substantial numbers of HIV-1 sequences covering the regions encoding both PR and RT, and representing a diversity of HIV-1 subtypes and sampling locations around the world (Arimide et al., 2018; Rasmussen et al., 2018; Wilhelmson et al., 2018). All resistance scores for all 1,006 sequences were completely concordant between the pipelines. Further details are provided in Ho, Ng, Renaud, & Poon (2018).

Performance

Performance was evaluated on a workstation running Ubuntu 18.04 LTS with an Intel Xeon E5-1650 v4 hexa-core CPU at 3.60 GHz and 16 GB of DDR4-2400 RAM with a gigabit network connection. *sierra-local* achieved mean [range] processing speeds of 47.08 sequences/second (seq/s) [45.07, 48.49] for PR, 16.20 seq/s [14.01, 19.97] for RT, and 14.99 seq/s [14.79, 15.56] for IN. A substantial fraction of processing time was consumed by subtyping. SierraPy, with the same dataset as previously described, yielded mean processing speeds of 16.01 seq/s [12.88, 17.60] for PR, 6.12 seq/s [4.83, 7.54] for RT, and 5.19 seq/s [5.05, 5.47] for IN. Although the size of sequence batches used in this performance comparison likely is a factor in the results by virtue of file writing and reading being done once per batch, the large batch size used minimizes the effect of these I/O processes on the overall runtime. Overall, *sierra-local* is able to process and return results for submitted query HIV-1 *pol* sequences roughly 3 times faster than SierraPy, depending on the nature of the sequences and the type of local computing resources available.

Concluding remarks

The distribution of the HIVdb resistance genotyping algorithm in a standardized format (ASI (Betts & Shafer, 2003)) is an important resource for HIV-1 research and clinical management, and an exemplary case of open science. *sierra-local* provides a convenient

framework to generate HIV drug resistance predictions from ASI releases in a secure environment and confers full control over data provenance. The ability to apply ASI-encoded algorithms locally (offline) also makes this part of the laboratory workflow robust to network availabilty may be particularly important for laboratories situated in resource-limited settings. We hope this lightweight, open-source implementation of the HIVdb ASI will further democratize HIV drug resistance genotyping across providers of HIV care.

Acknowledgements

We thank Philip Tzou for bringing NucAmino to our attention, and for his contributions to open science in the release of the Stanford HIVdb resistance program source code. This work was supported in part by the Government of Canada through Genome Canada and the Ontario Genomics Institute (OGI-131) and by a grant from the Canadian Institutes of Health Research (PJT-156178). The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

References

Arimide, D. A., Abebe, A., Kebede, Y., Adugna, F., Tilahun, T., Kassa, D., Assefa, Y., et al. (2018). HIV-genetic diversity and drug resistance transmission clusters in Gondar, Northern Ethiopia, 2003-2013. *PLoS ONE*, *13*(10), e0205446. doi:10.1371/journal.pone.0205446

Aziz, N., Zhao, Q., Bry, L., Driscoll, D. K., Funke, B., Gibson, J. S., Grody, W. W., et al. (2015). College of American Pathologists' laboratory standards for next-generation sequencing clinical tests. *Arch Pathol Lab Med*, *139*(4), 481–493. doi:10.5858/arpa.2014-0250-CP

Bernard, E. J., Azad, Y., Vandamme, A. M., Weait, M., & Geretti, A. M. (2007). HIV forensics: pitfalls and acceptable standards in the use of phylogenetic analysis as evidence in criminal investigations of HIV transmission. *HIV Med*, *8*(6), 382–387. doi:10.1111/j.1468-1293.2007.00486.x

Betts, B. J., & Shafer, R. W. (2003). Algorithm specification interface for human immunodefiency virus type 1 genotypic interpretation. *J Clin Microbiol*, *41*(6), 2792–2794. doi:10.1128/JCM.41.6.2792-2794.2003

Günthard, H. F., Aberg, J. A., Eron, J. J., Hoy, J. F., Telenti, A., Benson, C. A., Burger, D. M., et al. (2014). Antiretroviral Treatment of Adult HIV Infection: 2014 Recommendations of the International Antiviral Society-USA Panel. *JAMA*, *312*(4), 410–25. doi:10.1001/jama.2014.8722

Hart, S. A. S., Vardhanabhuti, S., Strobino, S. A., & Harrison, L. L. J. (2018). The impact of changes over time in the Stanford University Genotypic Resistance Interpretation algorithm. *J Acquir Immune Defic Syndr*, *79*(1), 1. doi:10.1097/QAI.0000000000001776

Ho, J. C., Ng, G. T., Renaud, M., & Poon, A. F. (2018). Sierra-local: A lightweight standalone application for secure HIV-1 drug resistance prediction. *bioRxiv*, 393207. doi:10.1101/393207

Liu, T. F., & Shafer, R. W. (2006). Web resources for HIV type 1 genotypic-resistance test interpretation. *Clin Infect Dis*, *42*(11), 1608–1618. doi:10.1086/503914

Mayer, K. H., Hanna, G. J., & D'Aquila, R. T. (2001). Clinical use of genotypic and phenotypic drug resistance testing to monitor antiretroviral chemotherapy. *Clin Infect Dis*, *32*(5), 774–782. doi:10.1086/319231

Meynard, J.-L., Vray, M., Morand-Joubert, L., Race, E., Descamps, D., Peytavin, G., Matheron, S., et al. (2002). Phenotypic or genotypic resistance testing for choosing antiretroviral therapy after treatment failure: A randomized trial. *AIDS*, *16*(5), 727–736.

Patil, H. K., & Seshadri, R. (2014). Big data security and privacy issues in healthcare. In *Big Data (BigData Congress), 2014 IEEE International Congress on* (pp. 762–765). IEEE. doi:10.1109/BigData.Congress.2014.112

Rasmussen, D. A., Wilkinson, E., Vandormael, A., Tanser, F., Pillay, D., Stadler, T., & Oliveira, T. de. (2018). Tracking external introductions of hiv using phylodynamics reveals a major source of infections in rural kwazulu-natal, south africa. *Virus Evolution*, *4*(2), vey037. doi:10.1093/ve/vey037

Shafer, R. W. (2006). Rationale and uses of a public HIV drug-resistance database. *J Infect Dis*, *194*(Supplement_1), S51–S58. doi:10.1086/505356

Tang, M. W., Liu, T. F., & Shafer, R. W. (2012). The HIVdb system for HIV-1 genotypic resistance interpretation. *Intervirology*, *55*(2), 98–101. doi:10.1159/000331998

Tural, C., Ruiz, L., Holtzer, C., Schapiro, J., Viciana, P., González, J., Domingo, P., et al. (2002). Clinical utility of HIV-1 genotyping and expert advice: The Havana trial. *AIDS*, *16*(2), 209–218.

Van Laethem, K., De Luca, A., Antinori, A., Cingolani, A., Perno, C. F., & Vandamme, A. M. (2002). A genotypic drug resistance interpretation algorithm that significantly predicts therapy response in HIV-1-infected patients. *Antivir Ther*, *7*(2), 123–129.

Wilhelmson, S., Månsson, F., Lindman, J. L., Biai, A., Esbjörnsson, J., Norrgren, H., Jansson, M., et al. (2018). Prevalence of HIV-1 pretreatment drug resistance among treatment naïve pregnant women in Bissau, Guinea Bissau. *PLoS ONE*, *13*(10), e0206406. doi:10.1371/journal.pone.0206406