BIOINFORMATICS APPLICATIONS NOTE
Vol. 30 no. 4 2014, pages 578–580
doi:10.1093/bioinformatics/btt724

Genetics and population analysis
Advance Access publication December 12, 2013

VIRAPOPS: a forward simulator dedicated to rapidly evolved viral populations

Michel Petitjean[1,2] and Anne Vanet[1,3,4,*]

[1]Department of Biology, Univ Paris Diderot, Sorbonne Paris Cité, [2]MTI, INSERM UMR-S 973, [3]CNRS, UMR7592, Institut Jacques Monod, F-75013 Paris and [4]Atelier de Bio Informatique, F-75005 Paris, France

Associate Editor: Jeffrey Barrett

ABSTRACT

Summary: Daily, mutability and recombination of RNA viruses result in the production of million variants. All these rapid genomic changes directly influence the functional sites of the protein, its 3D structure or its drug resistances. Therefore, it is important to simulate these drastic switches to determine their effects on virus populations. Many computer programs are able to simulate specific variations in DNA genomes, but are generally non-adapted to RNA viruses. They simulate site-specific selection pressures, but rarely pressures on covariant or on higher order correlated sites and no at all on synthetic lethal groups. That is why we felt it important to create VIRAPOPS, a forward simulator that models specific RNA virus functions. It was designed for computational biologists, biologists and virologists.

Availability and implementation: Free binaries are available through a software repository at http://petitjeanmichel.free.fr/itoweb.petitjean.freeware.html.

Contact: anne.vanet@univ-paris-diderot.fr

Received on October 9, 2013; revised on November 28, 2013; accepted on December 9, 2013

1 INTRODUCTION

RNA viruses have different characteristics from other populations. Their high mutability rate, 100 times higher than those of bacteria or eukaryotes, allows the production of daily million mutations. These multiple mutants, under selective pressure, permit the emergence of new viral variants. Several population simulators called 'forward' are able to impose a selective pressure on mutations: CDPOP (Landguth and Cushman, 2010), Nemo2.2 (Guillaume and Rougemont, 2006), SimuPop (Peng and Kimmel, 2005), SFS_Code (Hernandez, 2008) and Vortex (Lacy *et al*., 2009). These software solutions associate fitness to an allele and can combine several allele fitnesses. Unfortunately, this is done by adding or multiplying individual fitness that does not fit all situations. The covariation of two sites can define compensatory mutations (CM: the first mutation decreases the fitness of the organism and the second compensates the fitness lack of the first) or synthetic lethal (SL: mutations that are non-lethal when they are alone but become lethal when they are combined into a single genome). If pair of CMs are treated by existing simulator (Mostowy *et al*., 2011), SLs are not. In the SL case, both mutations taken separately does not change the fitness of the virus, it is only when they appear together that the fitness is changed. Then we cannot add or multiply fitness associated with each mutation or used the calculation done for CM. However, the synthetic lethals are used for therapy developments (Brouillet *et al*., 2010; Dahirel *et al*., 2011) and viral vaccines. Therefore, it becomes crucial to be able to simulate them. In addition, the higher order correlated mutations (involving covariation of CM or SL type, but with a number >2 sites) are not covered by conventional simulators.

Software programs dedicated to viral studies are able to predict drug resistances as HIVdb (human immunodeficiency virus database) program from Stanford university (Liu and Shafer, 2006), to store clinical data related to human immunodeficiency virus (HIV) and hepatitis C virus treatment and subtyping tools as REGA (Pineda-Pena *et al*., 2013) to calculate their pharmacokinetics and pharmacodynamics as Simcyp (Jamei *et al*., 2009), or to make epidemic model as FluTE (Chao *et al*., 2010), but do not allow large group of mutations simulation as described earlier in the text.

On one hand, Vortex, SimuPop, SFS_Code and CDPOP softwares dedicated to population genetics do not take into account the essential tools to treat high mutability of RNA viruses. On the other hand, software more directly dedicated to virus handle databases related to drugs, geographic spread or drug resistances (SimCYP, REGA, HIVdb). But none of these two types of family simulator is able to handle viral populations at high rates mutability. Yet it is now essential to use such tools to meet the ever greater emergence of new mutations including those related to drug resistances.

However, under selection pressure, new genomes appear quickly and increase the difficulty of treatment developments and stable vaccine design. In addition, due to high genome plasticity, the protein 3D structure becomes plural, making difficult docking work, and study of the active sites. Thus, it is necessary to simulate an RNA virus sequence population. We present in this manuscript an understandable, useful and easy-to-use RNA virus population simulator, designed to model a real situation such as virus population in a naïve or infected patient (Fig. 1).

2 OVERVIEW OF VIRAPOPS

2.1 Input

The input sequences must be in nucleotide Fasta format or text format (one sequence/line). The output format will be the same as the selected input format.

[*]To whom correspondence should be addressed.

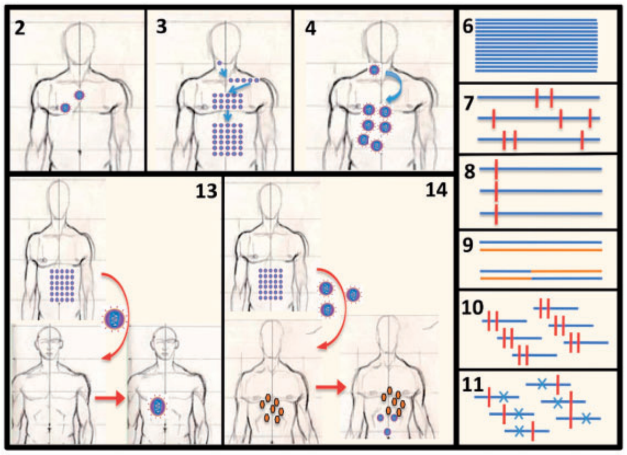

Fig. 1. Some VIRAPOPS options. (**2**) Sequences input. (**3**) Number of generations. (**4**) Number of budding virions per infection cycle. (**6**) Max number of sequences. (**7**) Polymerase mutation rate. (**8**) Hotspots and their mutation rates. (**9**) Recombination rate. (**10**) The CM groups and their mutation rate. (**11**) SL groups and their mutation rate. (**13**) Genetic drift (Naive patient infection). (**14**) Gene flow (surinfection)

2.1.1 Specific options related to RNA virus Large genomic changes induced by high mutation rate require to strengthen this pole. VIRAPOPS implements selective pressures at one or multiple sites, can force mutation in a particular amino acid and can generate CMs or SL pairs or groups. A complex multivariant network composed of CMs and SLs groups can also be generated.

2.1.2 Specific options related to the infection type The data may concern a seronegative patient infected by a single virus or more. In the case of surinfection of an already RNA virus-positive patient, an option of gene flow may be used. To deal with a cascade of infection cycle, a genetic drift option has been created. By selected generations on which genetic drifts and/or gene flows occurred, multiple cycles of infection and surinfection can be simulated on the same run. Outputs per generation are detailed tools for studying each infection step described previously.

2.1.3 Other options Extra options have been added to help biologists. It is possible to stop the simulator if an average percentage of mutations is reached. It is also possible to attain a maximum number of sequences and impose this number for future generations. Thus, the Hardy–Weinberg model can be followed if necessary. Hotspot option was created to deal with the fact that certain positions on certain genes are known to mutate more frequently than others (this is not due to selection pressure but because polymerase incorporates a wrong nucleotide at this place more easily than elsewhere). Finally, a redundancy option was created more specifically for computer biologists because for special purposes, they prefer to work on a unique sequence set (option available on DNA or protein sequences).

2.2 Ouput

Each generation produces a sequence set. The outputs are used to determine the impact of an infectious event occurring specifically in either generation. Its format shall be the same as the input. Regarding the HIV sequence, a Fasta output format will allow to determine the drug resistance status of each output sequences using the Sierra site at Stanford University (Liu and Shafer, 2006).

2.3 Example of use

2.3.1 HIV multitherapy and resistant mutation apparition Anti-HIV treatments are composed of several molecules against which resistance mutations can be selected. It could generate a conflict between a group of mutations needed to counteract drug A and another group of mutations needed to counteract the drug B. By simulating the appearance of mutations needed for both resistances, two scenarios can appear:

- The population will increase because of resistance mutation accumulation, showing no drug interaction issues. This result could reveal a potential future inactive treatment. The number of generations needed to get an inactive treatment may also be calculated using VIRAPOPS.
- The population remains undeveloped. Then, this therapy may be used because the resistance mutations against drugs A and B could not occur together.

To simulate this scenario, we propose to use a single sequence representing a virus that infects a seronegative patient. The mutation rate of the polymerase will be adjusted to 10^{-4}. The following variables must be entered in option 10: variability at each position, the covariability per positions couple both inferred from sequence alignment of untreated patients. Option 12 was created to enter selection pressures due to treatments (e.g. Nelfinavir D30N, I84A/V, N88S/D and L90M). The simulator will be launched on 50 generations, for example (option 3) with, if desired 60 budding virions/cycle (option 4). If the resistance mutations are viable, i.e. compatible with the variance and covariance that were previously entered, the sequence number will increase over time. Otherwise, the number of sequences will be stable or reduced.

2.3.2 Treatment interruption and consequences Some patients stop their treatment without informing their doctor (usually this latter noticing it when the patient's viral load increases again). These treatment stops can be simulated to evaluate what treatment would be most appropriate after such situation. This scenario will be coded in two stages. The first will be equivalent to that described in the previous paragraph. It will be the first part of the scenario in which the patient takes treatment. The batch of sequences generated will serve as input to simulate the cessation of treatment. In the second part, the same options are used with the exception of those corresponding to the treatment (Option 12). The new generated batch of sequences corresponds to the viral sequences of patient who has interrupted his/her treatment. It is important to note that the time of treatment interruption may be decreased or increased by changing the number of viral generations (option 3).

2.3.3 Using SL Oncology and virology are two areas where drugs that bind SLs are developed. Before the production stage, it would be advisable to see the possible interaction between these SL groups and necessary positions for the proper functioning of drugs already developed (especially in the case of HIV where multitherapies are generally necessary). So the docked drug on SL groups could avoid a resistance to another existing drug that would greatly strengthen the effect of this multiple therapy. To compute this simulation, two scenarios should be compared. The first one, similar to that of Section 2.3.1, simulates a patient with treatment. The second scenario retains the same options except those corresponding to the treatment selection pressure (this scenario is similar to the interruption processing part 2.3.2). Both simulations generate sets of sequences for which the SL couples will be determined (Brouillet *et al.*, 2010). The SL couples belonging to the 'treated patient' scenario and not to the 'untreated patients' scenario should be used as a therapeutic target for the development of new treatments.

2.3.4 Change treatment We would like to determine whether a patient treatment generates a resistance to a second treatment, which be administered to the same patient in addition to or instead of the first treatment. It would be inappropriate to change therapy to a new less effective one. So after programming the simulator as explained earlier in the text with the data for the current treatment, it is sufficient to determine whether the first treatment revealed mutations responsible for resistance to a future second treatment.

3 CONCLUSION

We developed VIRAPOPS, the first software dedicated to RNA virus population simulation. It allows virus variability representation, as variation site, covariation pair or higher order correlation in CM and SL. Moreover, it allows simulation of primary, secondary infections and cascade infections. It aims to be an easy-to-use tool and meet the most important needs of virologists and computational biologists.

ACKNOWLEDGEMENT

The authors thank P. Le Chien for his encouragements since 1993. This work is dedicated to Lou, 5-years old, who died of AIDS in 1997.

Conflict of Interest: none declared.

REFERENCES

Brouillet,S. *et al.* (2010) Co-lethality studied as an asset against viral drug escape: the HIV protease case. *Biol. Direct*, **5**, 40.

Chao,D.L. *et al.* (2010) FluTE, a publicly available stochastic influenza epidemic simulation model. *PLoS Comput. Biol.*, **6**, e1000656.

Dahirel,V. *et al.* (2011) Coordinate linkage of HIV evolution reveals regions of immunological vulnerability. *Proc. Natl Acad. Sci. USA*, **108**, 11530–11535.

Guillaume,F. and Rougemont,J. (2006) Nemo: an evolutionary and population genetics programming framework. *Bioinformatics*, **22**, 2556–2557.

Hernandez,R.D. (2008) A flexible forward simulator for populations subject to selection and demography. *Bioinformatics*, **24**, 2786–2787.

Jamei,M. *et al.* (2009) The Simcyp population-based ADME simulator. *Expert Opin. Drug Metab. Toxicol.*, **5**, 211–223.

Lacy,R.C. *et al.* (2009) Vortex: A stochastic simulation of the extinction process. Version 9.99. Chicago Zoological Society, Brookfield, Illinois, USA.

Landguth,E.L. and Cushman,S.A. (2010) cdpop: a spatially explicit cost distance population genetics program. *Mol. Ecol. Resour.*, **10**, 156–161.

Liu,T.F. and Shafer,R.W. (2006) Web resources for HIV type 1 genotypic-resistance test interpretation. *Clin. Infect. Dis.*, **42**, 1608–1618.

Mostowy,R. *et al.* (2011) The role of recombination for the coevolutionary dynamics of HIV and the immune response. *PLoS One*, **6**, e16052.

Peng,B. and Kimmel,M. (2005) simuPOP: a forward-time population genetics simulation environment. *Bioinformatics*, **21**, 3686–3687.

Pineda-Pena,A.C. *et al.* (2013) Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: performance evaluation of the new REGA version 3 and seven other tools. *Infect. Genet. Evol.*, **19**, 337–348.