BIOINFORMATICS

Vol. 00 no. 00 2015
Pages 1–2

An individual-based model forward simulator for transposable elements dynamics and evolution

Felipe Figueiredo [1,2]*, Claudio Struchiner [2]

[1]Programa de Biologia Computacional e Sistemas, Instituto Oswaldo Cruz (BCS/IOC/Fiocruz)
[2]Programa de Computao Cientfica, PROCC/Fiocruz

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Transposable Elements (TEs) are small genomic elements present in almost all genomes sequenced so far, that appear repeatedly in the individuals genome. Much debate is ongoing about their origin, but some mechanisms for the invasion of a host individual and the later spread within the population are known from both *in silico* and wet laboratory experiments.

In order to assess both the evolutionary forces that promote variation in TEs and their impact on host fitness, three independent levels of biological organization must be observed simultaneously: (a) the population demographics should be represented by a proper ecological model, (b) the TEs' population genetics should be given by a transposition model and (c) the TE sequences must evolve according to a known evolutionary model.

Results:

We present TRepid, an individual-based model that can be used to simulate TE invasion and fixation on an age-structured host population.

Host population dynamics features include random mating, various population growth models and recombination by crossing over of gametes. The TEs spread dynamics features include: transposition by either copy and paste or cut and paste, models that determine transposition rate, nucleotide substitution according to a selectable molecular evolutionary model, active and inactive TEs, inactivation of TE over generations, selective pressure on host individuals by accumulation of TEs and deleterious transposition events, among others.

Availability: The software is available under the GNU General Purpose Licence (GPL) version 3 from `https://launchpad.net/trepid`.

Contact: philsf79@gmail.com

1 INTRODUCTION

Several approaches have been proposed to understand and predict how the amount of TEs varies in hosts genomes (Le Rouzic and Deceliere, 2005; Hedges *et al.*, 2005; Struchiner *et al.*, 2005). Such mathematical models try to assess the invasion capabilities, accumulation and fixation of new copies and are usually based on mathematical and computational techniques.

Previous attempts in the modelling community focused in reproducing the dynamics predicte by differential or difference equations that express how the total amount of TE copies vary in terms of acquisition of new copies and excision of old ones (Quesneville and Anxolabehere, 1997, 1998; Deceliere *et al.*, 2005, 2006).These types of models can typically be characterized in a continuum between two paradigm extremes: the *master gene* model, in which only one TE is active and generates inactive copies, and the *transposon* model in which every TE actively produces active copies, leading to an exponential growth of TE copy number in the absence of some regulation mechanism (Katzourakis *et al.*, 2005). Phylogenetic analysis can then provide the means to asses where in this spectrum a given TE family resides (Brookfield and Johnson, 2006; Johnson and Brookfield, 2006).

Most of the literature on models describing TE dynamics suffer from the lack of empirical data against which these models could be tested. It means that model diagnosis, an important step in model development, is missing. By looking at the sequences of families of transposable elements from the same genome, one can make inferences about their phylogenetic tree. This tree is influenced by, among other things, the changes in the number of elements of a given family over evolutionary time. Thus, in principle, one can make inferences about changes in the number of mobile elements in the genome from the phylogeny of the elements. Therefore, it becomes clear that examining the population dynamics of TEs through coalescent approaches is going to be highly informative (Brookfield, 2005; Brookfield and Johnson, 2006; Struchiner *et al.*, 2009). The main contribution of our work is to devise a simulation framework that mimics the empirical sequence data on TE dynamics where this dynamics is known. By exploring this simulation framework, we hope to validate the use of colescent models to estimate potential parameters, and associated uncertainty in the estimation process, that describe TE invasion. In doing so, we hope to contribute an important tool to the debate about the introduction of transposable elements as part of genetic drive systems moving genes through disease vector populations.

2 METHODS

The simulator takes into consideration distinct and independent modeling techniques for each level of biological organization: a population model, a transposition model and an evolutionary model of the DNA sequences. Each

*to whom correspondence should be addressed

of these sections are based on either differential or difference equations, or probabilistic models.

At the ecological level the population model produces a trend for the population dynamics. Currently there are models for constant and exponential growth, and growth with saturation (logistic and Hassel equations, see SOM for details. At the start of each generation, the necessary ammount of sexually mature individuals is sampled and coupled to generate the required ammount of offspring to follow the ecological model as closely as possible, notwithstanding fitness effects from TEs. The modular framework provide the means to implement any other ecological model. An age structure is also optionally available in discrete age classes.

At the population genetics level, the transposition model does the same thing for the ammount of TEs in each new individual, based on the ammount of TEs in the parental gametes. Transposition models are available with selection impact (Struchiner *et al.*, 2005) and without (Le Rouzic and Deceliere, 2005).

At the sequence level the evolutionary model determines how the TE sequences change over time, after sucessive transposition events, and an aging structure for TEs that escape excision from the host chromosomes.

2.1 The generation algorithm

The algorithm that happens at each generation models the basic life cycle of a diploid sexual species subjected to transposition events during gametogenesis.

1. Host couples are chosen randomly from available mature hosts at the beginning of each reproductive season. Males and females are chosen with and without replacement, respectively.
2. Each adult bears new gametes after transposition and recombination.
3. Transposition draws a recruitment amount of new TE copies and the deletion amount of excised copies from the transposition model. This changes the content of the gametes in terms of availability of TEs.
4. Mutations are sampled from the evolutionary model for newly created TE copies. If “cut and paste” transposition is being used, sample mutations for the original copy also. The same does not happen for “copy and paste”.
5. Recombination provides additional shuffling of gamete contents.
6. Mutations are sampled from the evolutionary model for all existing TE copies.
7. Each couple gives birth to a number of offspring defined by the user as a parameter.
8. The fitness cost from TEs in newborn individuals is calculated and any that exceeds a given threshold is killed before birth and removed from the population.
9. The age of every surviving individual is incremented at the end of the generation.

3 CONCLUSION

In this article we describe a computational model composed of a forward-time individual-based model for the population genetics and molecular evolution of transposable elements.

Population genetics software exist for both forward simulations (Carvajal-Rodriguez, 2008; Guillaume and Rougemont, 2006; Peng and Kimmel, 2005; Padhukasahasram *et al.*, 2008; Hernandez, 2008) and backward-time simulations (Hudson, 2002; Teshima and Innan, 2009), although most of them simply count the distribution and availability of a set of alelles that populate a given *locus* or *loci*. Similarly, there are simulators for transposition phenomena in host populations (Deceliere *et al.*, 2006) but they assume that TEs don’t change over time. As far as we are aware there is no simulator dealing with all three levels of biological organization concomitantly as well as considering how one level affect each other.

Additionally, at the end of each simulation an individual is optionally sampled from the population so an additional level of ecological modeling is being implicitly considered. This provides the means to take into account a sampling distribution in a statistical ecology framework.

ACKNOWLEDGEMENT

Funding: This work was partially supported by a Bill & Melinda Gates Foundation grant (FIXME), and a PhD scholarship by CAPES (FIXME).

REFERENCES

Brookfield, J. F. (2005). Evolutionary forces generating sequence homogeneity and heterogeneity within retrotransposon families. *Cytogenet Genome Res*, **110**(1-4), 383–91.

Brookfield, J. F. and Johnson, L. J. (2006). The evolution of mobile DNAs: when will transposons create phylogenies that look as if there is a master gene? *Genetics*, **173**(2), 1115–23.

Carvajal-Rodriguez, A. (2008). GENOMEPOP: a program to simulate genomes in populations. *BMC Bioinformatics*, **9**, 223.

Deceliere, G., Charles, S., and Biemont, C. (2005). The dynamics of transposable elements in structured populations. *Genetics*, **169**(1), 467–74.

Deceliere, G., Letrillard, Y., Charles, S., and Biemont, C. (2006). TESD: a transposable element dynamics simulation environment. *Bioinformatics*, **22**(21), 2702–3.

Guillaume, F. and Rougemont, J. (2006). Nemo: an evolutionary and population genetics programming framework. *Bioinformatics*, **22**(20), 2556–7.

Hedges, D. J., Cordaux, R., Xing, J., Witherspoon, D. J., Rogers, A. R., Jorde, L. B., and Batzer, M. A. (2005). Modeling the amplification dynamics of human Alu retrotransposons. *PLoS Comput Biol*, **1**(4), e44.

Hernandez, R. D. (2008). A flexible forward simulator for populations subject to selection and demography. *Bioinformatics*, **24**(23), 2786–7.

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**(2), 337–8.

Johnson, L. J. and Brookfield, J. F. (2006). A test of the master gene hypothesis for interspersed repetitive DNA sequences. *Mol Biol Evol*, **23**(2), 235–9.

Katzourakis, A., Rambaut, A., and Pybus, O. G. (2005). The evolutionary dynamics of endogenous retroviruses. *Trends Microbiol*, **13**(10), 463–8.

Le Rouzic, A. and Deceliere, G. (2005). Models of the population genetics of transposable elements. *Genet Res*, **85**(3), 171–81.

Padhukasahasram, B., Marjoram, P., Wall, J. D., Bustamante, C. D., and Nordborg, M. (2008). Exploring population genetic models with recombination using efficient forward-time simulations. *Genetics*, **178**(4), 2417–27.

Peng, B. and Kimmel, M. (2005). simuPOP: a forward-time population genetics simulation environment. *Bioinformatics*, **21**(18), 3686–7.

Quesneville, H. and Anxolabehere, D. (1997). A simulation of P element horizontal transfer in Drosophila. *Genetica*, **100**(1-3), 295–307.

Quesneville, H. and Anxolabehere, D. (1998). Dynamics of transposable elements in metapopulations: a model of P element invasion in Drosophila. *Theor Popul Biol*, **54**(2), 175–93.

Struchiner, C. J., Kidwell, M. G., and Ribeiro, J. M. C. (2005). Population Dynamics of Transposable Elements: Copy Number Regulation and Species Invasion Requirements. *Journal of Biological Systems*, **13**(4), 455–475.

Struchiner, C. J., Massad, E., Tu, Z., and Ribeiro, J. M. (2009). The tempo and mode of evolution of transposable elements as revealed by molecular phylogenies reconstructed from mosquito genomes. *Evolution*, **63**(12), 3136–46.

Teshima, K. M. and Innan, H. (2009). mbs: modifying Hudson’s ms software to generate samples of DNA sequences with a biallelic site under selection. *BMC Bioinformatics*, **10**, 166.