In silico experimental evolution provides independent and challenging benchmarks for comparative genomics

Priscila Biller[1], Éric Tannier[2,3], Guillaume Beslon[3,4], Carole Knibbe *[3,5]

[1] University of Campinas [Campinas] (UNICAMP) – São Paulo, Brésil
[2] Laboratoire de Biométrie et Biologie Évolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I, INRIA – 43 boulevard du 11 Novembre 1918, F-69 622 Villeurbanne Cedex, France
[3] BEAGLE (Insa Lyon / INRIA Grenoble Rhône-Alpes / UCBL) – INRIA, Institut National des Sciences Appliquées [INSA] - Lyon, Université Claude Bernard - Lyon I (UCBL) – Antenne INRIA Lyon la Doua Bâtiment CEI-1, 66 boulevard Niels Bohr, F-69 603 Villeurbanne, France
[4] Laboratoire d'InfoRmatique en Image et Systèmes d'information (LIRIS) – Institut National des Sciences Appliquées [INSA], CNRS : UMR5205 – France
[5] Laboratoire d'InfoRmatique en Image et Systèmes d'information (LIRIS) – Université Claude Bernard - Lyon I (UCBL), CNRS : UMR5205 – France

Session génomique des populations
mercredi 29 11h10
Amphi Mérieux

The following extended abstract is a highlight of [7].

A common concern in all evolutionary studies is the validity of the methods and results. Results relate to events that were supposed to occur in a deep past (up to 4 billion years) and they have no other trace today than the present molecules used by comparative methods.

As we cannot travel back in time to verify the results, there are several ways to assess the validity of molecular evolution studies: theoretical considerations about the models and methods (realism, consistency, computational complexity, robustness, model testing, ability to generate a statistical support or a variety of the solutions) [23], coherence with fossil records [25], or ancient DNA [11], or empirical tests when the solution is known, on experimental evolution [17] or simulations. Each method has its caveats. Models for inference have to adopt a compromise between realism, consistency and complexity. Ancient DNA is rarely available, usually not in an assembled shape. Fossils are also rare and provide a biased sampling of ancient diversity. Experimental evolution is expensive, time- consuming and limited in the number of generations it can provide.

Simulation is the most popular validation tool. Genome evolution can be simulated in silico for a much higher number of generations than in experimental evolution, much faster and at a lower cost. All the history can be recorded in details, and compared with the inference results. A problem with simulations, however, is that they necessarily oversimplify genome evolution processes. Moreover, very often, even if they are designed to be used by another team for inference [4, 15, 14, 10, 22], they encode the same simplifications as the inference methods. For example, only fixed mutations are generated because only these are visible by inference methods; selection is tuned to fit what is visible by the inference methods; genes are often evolutionary units in simulations because they are the units taken for inference. Everything is designed thinking of the possibilities of the inference methods.

This mode of ad-hoc simulation has been widely applied to test estimators of rearrangement distances, and in particular inversion distances [9, 12, 5, 21, 6]. The problem consists in comparing two genomes and estimating the number of inversions (a rearrangement that reverses the reading direction of a genomic segment) that have occurred in the evolutionary lineages separating them. To construct a solution, conserved genes or synteny blocks are detected in the two genomes, and a number of inversions explaining the differences in gene orders is estimated. A lot of work has

*. Intervenant

consisted in finding shortest scenarios [13]. Statistical estimations need a model. The standard and most used model depicts genomes as permutations of genes and assumes that an inversion reverses a segment of the permutation, taken uniformly at random over all segments. When simulators are designed to validate the estimators, they also use permutations as models of gene orders, and inversions on segments of this permutations, chosen uniformly at random. Estimators show good performances on such simulations, but transforming a genome into a permutation of genes is such a simplification from both parts that it means nothing about any ability to estimate a rearrangement distance in biological data [8].

We propose to use simulations that were not designed for validation purposes. It is the case, in artificial life, of in silico experimental evolution [18], and in particular of the Aevol platform [19, 3]. Aevol contains, among many other features, all what is needed to test rearrangement inference methods. The genomes have gene sequences and non coding sequences organized in a chromosome, and evolve with inversions, in addition to substitutions, indels, duplications, losses, translocations. Rearrangements are chosen with a uniform random model on the genome, which should fit the goals of the statistical estimators, but is different from a uniform random model on permutations [8].

We tested 10 different estimators of inversion distance found in the literature, one shortest path estimator and 9 statistical estimators on 18 different datasets generated by Aevol. The difference with ad-hoc simulations is striking. Whereas good results were largely reported for ad-hoc simulations, most estimators completely fail to give a close estimate in a vast majority of conditions. As soon as the true number of events exceeds about n/3 (where n is the number of genes), most estimators significantly underestimate the true value. This highly contrasts with the claimed performances of these estimators. For example the shortest path estimator is supposed to have great chance of giving the right value up to n/2 [16], while all statistical estimators have been tested on simulations and reported to give the right value far above n [9, 20, 12, 5, 21, 2, 6, 8].

We argue, based on the differences in performances of some estimators, that our datasets are not artefactually difficult (nor purposely made difficult), and that the poor results encountered here are susceptible to reflect real results on biological data. Indeed part of the failure of the estimators can be explained by this ignorance of intergene sizes, because the only one handling intergene sizes performs significantly better. We investigated this further in [8].

Part of the discrepancy between the true value and the estimated value still remains unexplained. The complexity of the real scenarios probably blurs the signal that estimators are able to capture. But again, this complexity is not a specificity of Aevol, and is probably encountered in biological data. So by this simple experiment we can worry that none of the existing estimators of rearrangement distance would be able to produce a plausible value on real genomes.

We tested only the estimation of the number of inversions. But only with the runs we have already computed, a lot more can be done: estimation of the proportion of translocations (transposition of a block of DNA at an other locus) as in [1], or estimating both inversions and duplications as in [24]. For the moment the sequences are made of 0s and 1s, which is not a problem to study gene order, but can be disturbing for sequence analyses. This way of coding sequences is on another hand a good sign that Aevol was not developed for benchmarking purposes. In a close future, nucleotidic and proteic sequences with the biological alphabet will be added to extend the benchmarking possibilities of the model.

Also we worked with only one lineage, and compare only two genomes here (final versus ancestral), because Aevol currently evolves only one population at a time. A useful addition will be speciation processes, in order to be able to compare several genomes.

As a final note, we would like to point out the singular kind of interdisciplinarity experimented in this study. Obviously communities from comparative genomics and artificial life have to work together in order to make such results possible. But, on the opposite, these results are only possible because both communities first worked in relative isolation. If they had defined their

working plans together, spoke to each other too often or influenced each other's way of thinking evolutionary biology, the work would have lost some value. Indeed, what makes the difficulty here for comparative genomicists is that they have to infer histories on data for which they have no stranglehold on the processes, just as for biological data, but on which they also have the correct answer, just not as for biological data.

References

[1] N Alexeev, R Aidagulov, and MA Alekseyev. A computational method for the rate estimation of evolutionary transpositions. *Bioinformatics and Biomedical Engineering*, pages 471–480, 2015.

[2] N Alexeev and Max A. Alekseyev. Estimation of the true evolutionary distance under the fragile breakage model. *Arxiv*, 2015.

[3] Berenice Batut, David P. Parsons, Stephan Fischer, Guillaume Beslon, and Carole Knibbe. In silico experimental evolution: a tool to test evolutionary scenarios. *BMC Bioinformatics*, 14 (S15):S11, 2013.

[4] R G Beiko and R L Charlebois. A simulation test bed for hypotheses of genome evolution. *Bioinformatics*, 23(7):825–831, April 2007.

[5] Nathanal Berestycki and Rick Durrett. A phase transition in the random transpo- sition random walk. *Probability Theory and Related Fields*, 136:203–233, 2006.

[6] Priscila Biller, Laurent Guéguen, and Éric Tannier. Moments of genome evolution by double cut-and-join. *BMC Bioinformatics*, 16, 2015.

[7] Priscila Biller, Carole Knibbe, Guillaume Beslon, and Éric Tannier. Comparative genomics on artificial life. In *Proceedings of Computability in Europe (CiE) 2016, LNCS*. Springer, 2016.

[8] Priscila Biller, Carole Knibbe, Laurent Guéguen, and Éric Tannier. Breaking good: accounting for the diversity of fragile regions for estimating rearrangement distances. *Genome Biology and Evolution*, 2016, in press.

[9] Alberto Caprara and Giuseppe Lancia. Experimental and statistical analysis of sorting by reversals. In David Sankoff and Joseph H. Nadeau, editors, *Comparative Genomics*, pages 171–183. Springer, 2000.

[10] Daniel A. Dalquen, Maria Anisimova, Gaston H. Gonnet, and Christophe Dessimoz. ALF – a simulation framework for genome evolution. *Mol Biol Evol*, 29(4):1115–1123, Apr 2012.

[11] Wandrille Duchemin, Vincent Daubin, and Éric Tannier. Reconstruction of an ancestral *Yersinia pestis* genome and comparison with an ancient sequence. *BMC Genomics*, 16 Suppl 10:S9, 2015.

[12] Niklas Eriksen and Axel Hultman. Estimating the expected reversal distance after a fixed number of reversals. *Advances in Applied Mathematics*, 2004.

[13] G. Fertin, A. Labarre, I. Rusu, É. Tannier, and S. Vialette. *Combinatorics of Genome Rearrangements*. MIT press, London, 2009.

[14] William Fletcher and Ziheng Yang. Indelible: a flexible simulator of biological sequence evolution. *Mol Biol Evol*, 26(8):1879–1888, Aug 2009.

[15] B G Hall. Simulating DNA Coding Sequence Evolution with EvolveAGene 3. *Molecular Biology and Evolution*, 25(4):688–695, February 2008.

[16] Hannenhalli and P. A. Pevzner. Transforming men into mice (polynomial algorithm for genomic distance problem). In *Foundations of Computer Science, 1995*. Proceedings., 36th Annual Symposium, 1995.

[17] D. M. Hillis, J. J. Bull, M. E. White, M. R. Badgett and I. J. Molineux. Experimental phylogenetics: generation of a known phylogeny. *Science*, 255(5044):589–592, Jan 1992.

[18] Thomas Hindré, Carole Knibbe, Guillaume Beslon, and Dominique Schneider. New insights into bacterial adaptation through in vivo and in silico experimental evolution. *Nature Reviews Microbiology*, 10:352–365, May 2012.

[19] Carole Knibbe, Antoine Coulon, Olivier Mazet, Jean-Michel Fayard, and Guillaume Beslon. A long-term evolutionary pressure on the amount of noncoding DNA. *Molecular Biology and Evolution*, 24(10):2344–2353, Oct 2007.

[20] B. Larget, D. L. Simon, and J.B. Kadane. On a bayesian approach to phyloge- netic inference from animal mitochondrial genome arrangements (with discussion). *Journal of the Royal Statistical Society, B*, 64:681–693, 2002.

[21] Yu Lin and Bernard M E. Moret. Estimating true evolutionary distances under the DCJ model. *Bioinformatics*, 24(13):i114–i122, Jul 2008.

[22] Diego Mallo, Leonardo De Oliveira Martins, and David Posada. Simphy: Phyloge- nomic simulation of gene, locus, and species trees. *Syst Biol*, Nov 2015.

[23] M. Steel and D. Penny. Parsimony, likelihood, and the role of models in molecular phylogenetics. *Mol Biol Evol*, 17(6):839–850, Jun 2000.

[24] KM Swenson, M Marron, JV Earnest-DeYoung, and BME Moret. Approximating the true evolutionary distance between two genomes. *Journal of Experimental Algorithmics*, 12, 2008.

[25] Gergely J. Szollosi, Bastien Boussau, Sophie S. Abby, Éric Tannier, and Vincent Daubin. Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proc Natl Acad Sci U S A*, 109(43):17513–17518, Oct 2012.

Mots clefs : simulation, genome evolution, inversion distance, intrachromosomal rearrangements, benchmark, individual based modeling, comparative genomics