A 4-base model for the Aevol *in-silico* experimental evolution platform

Vincent Liard, Jonathan Rouzaud-Cornabas, Nicolas Comte, Guillaume Beslon

Université de Lyon, CNRS, Inria Beagle team, INSA-Lyon, LIRIS, UMR5205, F-69621, France
vincent.liard@inria.fr

Abstract

In this paper, we describe a new digital genetics model based on the Aevol artificial life simulator. Aevol is a computational platform designed to study populations of digital organisms evolving under various conditions. It has been extended in two directions. First, we have extended the genomic code from a binary one to a 4-base one, allowing for more realistic genomic sequence and eases the usage of Aevol as a benchmarking tool for comparative genomics. Second, we have replaced the Aevol continuous phenotype representation by a discrete one inspired by Fisher's Geometric Model. By doing so, we will be able to validate Aevol results against population genetics theory.

Why a new model?

There is a twofold motivation for extending the regular Aevol model: benchmarking phylogenetic algorithms and embedding Fisher's Geometric Model of evolution (FGM).

Benchmarking

Molecular evolutionary methods and tools are difficult to validate, as we have almost no direct access to ancient molecules. In Alife platforms such as Avida or Aevol, phylogenies are exactly recorded. The final population resulting from such *in silico* experiments can be analyzed by the phylogenetic algorithms to recover the phylogenetic tree. This process makes it possible to compare the trees *inferred* by these algorithms to the *actual* tree that was recorded along the way of artificial evolution.

This approach has recently been applied to test various estimators of inversion distance (Biller et al., 2016b), revealing their limits and suggesting important improvement directions (Biller et al., 2016a). However, Aevol uses a binary representation for the genomic sequence, thus strongly limiting its usability as a benchmarking tool. This limitation called for a new model based on 4-nucleotide sequences.

Fisher's Geometric Model

The other intent of this new model is to enable a direct comparison of Aevol results in terms of population genetics and, more precisely, in terms of FGM. Indeed, one of the drawbacks of digital genetics and artificial life models is their difficulty to crosstalk with other theoretical approaches in evolutionary biology. FGM is a simple mathematical model describing the qualitative behavior of evolution (Fisher, 1930; Tenaillon, 2014). Assessing compatibility between Aevol's model and FGM will make it possible to validate Aevol predictions in cases where FGM alone provides a clear notion of what is expected from evolution.

Aevol-ACGT model

In Aevol, a population of individuals evolves through a classical mutation-selection process. The specificity of Aevol lies in the genotype-to-phenotype mapping that finely models what is observed in bacteria. A circular double-stranded DNA sequence is transcribed into a set of mRNAs. These mRNAs are then parsed in search for Coding DNA Sequences (the "genes") that are translated into proteins through an artificial genetic code. Finally, the proteins are combined to compute the individual's phenotype. We refer the reader to previously published work for a complete description of the binary model and the results obtained so far (Knibbe et al., 2007; Batut et al., 2013; Misevic et al., 2015).

As in the classical Aevol, in Aevol-ACGT the digital organisms own a sequence of nucleotides genotype that encodes for a mathematical phenotype. The fitness of an organism is then compared with a predefined *phenotypic target* and the distance between the encoded phenotype and the target is used to compute the fitness. However, in the new model the genotype is a sequence on a 4-character alphabet (equivalent to ACGT) while the phenotype is modeled by a set of continuous traits (as in FGM). The phenotypic target defines the optimal value for all the traits under selection. Then the fitness w is computed from the distance between the phenotype and the phenotypic target through the classical Gaussian-based function of FGM:

$$w = e^{-\frac{1}{2}\sum_{i=1}^{n}(z_i - Z_i)^2} \quad (1)$$

Carole Knibbe et al, eds., Proceedings of the ECAL 2017, Lyon, France, 4-8 September 2017,
(Cambridge, MA: The MIT Press, ©2017 Massachusetts Institute of Technology).

where n is the number of traits under selection (corresponding to the complexity of the phenotype in FGM), z_i is the value of the ith trait and Z_i is its target value.

The genotype-to-phenotype map

One of the key properties of the Aevol model is its genotype-to-phenotype mapping that uses a four level process (DNA-RNA-Protein-Phenotype) akin to the classical central dogma of molecular biology. In Aevol-ACGT we use a similar mapping. However, the transition from a 2-base code to a 4-base one increases complexity.

In Aevol, we used 2 bases at the DNA and RNA levels and 6 "amino-acids" (AA) to describe the proteins (*i.e.* 8 codons minus the start and the stop ones). This enabled us to decode the protein sequence using 3 binary variables ($6 = 2 \times 3$). Hence, in Aevol the mathematical phenotype is modeled as a sum of 3-parameter functions (called *kernel*).

In Aevol-ACGT, the 4-base DNA sequence is translated into amino-acid sequences using the (degenerate) canonical genetic code. Using the same encoding as in Aevol would thus lead to 10-parameter kernel which could be difficult to calibrate in practice. Moreover, the combinatorics of the (arbitrary) AA-parameters association could be problematic. To overcome this difficulty, we propose to encode the parameters of the kernel using non-binary codes: the 20 AA are grouped into classes and all the AA of a same class are used to compute a same parameter. Multiple AA classifications have been proposed in the literature based on different criteria. Our model uses the classification proposed in Solis, 2015 that clusters amino-acids into 6 classes. Hence, our kernel has 6 parameters (one per AA class, encoded though a n-ary code depending on the size n of the class): the phenotypic space is a 2D space and each protein contributes to the phenotype by adding a 2D Gaussian to the phenotype described by 4 parameters: x, y, σ and h, the two remaining parameters describing the epistatic property of the protein following the Hansen and Wagner (2001) multilinear model. Finally, the n traits under selection are n points randomly or regularly scattered over the 2D phenotypic space and for which the target value Z_i is specified and can be compared to the phenotype value at the same position z_i (equation 1).

Evolutionary loop

The core of Aevol has not changed since its introduction in Knibbe et al., 2007. It consists of a loop describing the cycle of generations. All the individuals of a given population are synchronized in the sense that they all live and die in the time frame of their generation.

At first, an initial phase gives birth to a population of clones of a single viable individual. This is accomplished by drawing random DNAs until one happens to show a fitness better than 0. The evolutionary loop can be described as follows: (1) each organism goes through the transcription-translation process which ultimately yields a measure of its fitness related to the environment. (2) The fitnesses of all the organisms are then compared to decide how much offspring each individual will have. (3) Each offspring undergoes random DNA mutations (single base mutation, insertion or deletion of up to six nucleotides, duplications, deletions, translocations, inversions). (4) The loop iterates to the following generation by going back to step 1.

Conclusion

Aevol-ACGT has been developed in order to enable a better communication between digital genetics and artificial life and comparative genomics on the one hand and population genetics on the other hand. In this paper, we summarized how we modified the Aevol model to develop Aevol-ACGT.

We are currently in the process of calibrating and validating the Aevol-ACGT model. The first results show that the model is able to evolve complex organisms owning hundreds of genes. At the end of this calibrating phase, we will be able to use this new model to simulate complex evolutionary scenarios and to propose them as benchmarks to test phylogenomics tools. We will also be able to test, in the model, some of the theoretical properties that have been identified in FGM (such as the "cost of complexity") and to link population genetics theory with artificial life simulations.

References

Batut, B., Parsons, D. P., Fischer, S., Beslon, G., and Knibbe, C. (2013). In silico experimental evolution: a tool to test evolutionary scenarios. *BMC Bioinformatics*, 14(15):S11.

Biller, P., Guéguen, L., Knibbe, C., and Tannier, E. (2016a). Breaking Good: Accounting for Fragility of Genomic Regions in Rearrangement Distance Estimation. *Genome Biology and Evolution*, 8(5):1427–1439.

Biller, P., Knibbe, C., Beslon, G., and Tannier, E. (2016b). Comparative genomics on artificial life. In *Conference on Computability in Europe*, pages 35–44.

Fisher, R. A. (1930). The genetical theory of natural selection – a complete variorum edition. *Oxford University Press*.

Hansen, T. and Wagner, G. (2001). Model genetic architecture : a multilinear theory of gene interaction. *Theoretical Population Biology*, 59:61–86.

Knibbe, C., Coulon, A., Mazet, O., Fayard, J.-M., and Beslon, G. (2007). A long-term evolutionary pressure on the amount of noncoding DNA. *Molecular Biology and Evolution*, 24(10):2344–2353.

Misevic, D., Frénoy, A., Lindner, A. B., and Taddei, F. (2015). Shape matters: Lifecycle of cooperative patches promotes cooperation in bulky populations. *Evolution*, 69(3):788–802.

Solis, A. D. (2015). Amino acid alphabet reduction preserves fold information contained in contact iinteraction in proteins. *Proteins*, 83(12):2198–2216.

Tenaillon, O. (2014). The utility of Fisher's geometric model in evolutionary genetics. *Annual Review of Ecology, Evolution, and Systematics*, 45:179–201.

Carole Knibbe et al, eds., Proceedings of the ECAL 2017, Lyon, France, 4-8 September 2017,
(Cambridge, MA: The MIT Press, ©2017 Massachusetts Institute of Technology).