Mini-symposium JOBIM 2022

Simulation of biological sequences and evolutionary processes

Program

15h00 – 15h15	Introduction to the mini-symposium
15h15 – 15h45	**Catherine Matias** – Cophylogeny reconstruction in host-parasites systems via Approximate Bayesian Computation and allowing for multiple associations
15h45 – 16h15	**Burak Yelmen** – Deep convolutional and conditional neural networks for large-scale genomic data generation
16h15 – 16h45	Coffee break
16h45 – 17h15	**Guillaume Beslon** – Simulate genomic sequences evolution along complex phylogenetic trees with Aevol
17h15 – 17h45	**William Boulton** – Simulating SARS-CoV-2 sequences: phylogenies and metagenomes
17h45 – 18h00	Round table and discussions

Abstracts

Catherine Matias (Laboratoire de Probabilités, Statistique et Modélisation, UMR CNRS 8001, Sorbonne Université, Université de Paris, France)

Cophylogeny reconstruction in host-parasites systems via Approximate Bayesian Computation and allowing for multiple associations

Nowadays, the most used method in studies of the coevolution of hosts and symbionts is phylogenetic tree reconciliation. Most algorithms for host–parasite reconciliation use an event-based model, where the events include in general (a subset of) cospeciation, duplication, loss, and host switch. All known parsimonious event-based methods then assign a cost to each type of event in order to find a reconstruction of minimum cost. A crucial issue is that from a biological point of view, reasonable cost values for an event-based parsimonious reconciliation are not easily chosen, and this choice strongly influences the reconciliation obtained. Thus, different approaches have been developed to infer such cost values for a given pair of host and symbiont trees.

We developed an algorithm, called Coala, for estimating the frequency of the events based on an approximate Bayesian computation (ABC) approach. The benefits of this method are 2-fold: i) it provides more confidence in the set of costs to be used in a reconciliation; and ii) it allows estimation of the frequency of the events in cases where the data set consists of trees with a large number of taxa.

However, a major limitation of all these approaches (including Coala) is their inability to model the 'invasion' of different host species by a same symbiont species (referred to as a spread event), which is thought to happen in symbiotic relations.

To mention one example, the same species of insects may pollinate different species of plants. This results in multiple associations observed between the symbionts and their hosts (meaning that a symbiont is no longer specific to a host), that are not compatible with the current methods of coevolution.

After presenting the context and the tool Coala, I will discuss an extension called ACoala, which for a given pair of host and symbiont trees, estimates the probabilities of the cophylogeny events, in presence of spread events, still relying on an ABC approach.

The algorithm that we propose, by including spread events, enables the multiple associations to be taken into account in a more accurate way, inducing more confidence in the estimated sets of costs and thus in the reconciliation of a given pair of host and symbiont trees. Its rooting in the tool Coala allows it to estimate the probabilities of the events even in the case of large datasets. We will illustrate the performance of the method on real host-symbionts associations datasets.

Burak Yelmen (Laboratoire de Recherche en Informatique, UMR CNRS 8623, Université Paris-Sud, Université Paris-Saclay, Paris, France)

Deep convolutional and conditional neural networks for large-scale genomic data generation

Applications of generative models for genomic data have gained significant momentum in the past few years, with scopes ranging from data characterization to generation of simulated genomes. Genomic data is structured and a variety of processes such as recombination, mutations, natural selection and past demographic events contribute to the complexity of genomic data structure. In our previous study, we demonstrated that generative adversarial networks (GANs) and restricted Boltzmann machines (RBMs) can be used to create novel high-quality artificial genomes which can preserve the complex characteristics of real genomes such as population structure, linkage disequilibrium and selection signals. However, a major drawback of these models is scalability, as the large feature space of genome-wide data increases computational complexity vastly.

To address this issue, we explored different convolutional GAN and variational autoencoder (VAE) architectures along with a new conditional RBM framework for generating artificial genomes with high SNP coverage. All these networks implicitly learn the varying landscape of recombination in order to capture complex correlation patterns along the genome and generate a wide diversity of plausible haplotypes. We performed comparative analyses to assess both the quality of these generated haplotypes and the amount of privacy leakage from the training data. As the importance of genetic privacy becomes more prevalent, the need for effective privacy measures for genomic data increases. In the near future, large-scale artificial genome databases can be assembled to provide easily accessible surrogates of real databases, allowing researchers to perform studies with diverse genomic data within a safe ethical framework.

Guillaume Beslon (Projet INRIA BEAGLE – Artificial Evolution and Computational Biology, INRIA Antenne Lyon la Doua, Villeurbanne, France)

Simulate genomic sequences evolution along complex phylogenetic trees with Aevol

Simulating the evolution of a biological sequence requires the integration of mutation, selection and genetic drift processes. However, the coupling between these three processes is based on the possibility of attributing a reproductive capacity to any sequence likely to appear – by mutation – within the population. Several approaches have thus been proposed, including "artificial chemistry" which define an abstract algorithmic process for decoding the genome and simulate evolution in the virtual biochemical world generated. However, these approaches generate sequences that have little in common with real biological sequences: thus, Avida, one of the best-known software in this field, simulates sequences composed of assembler pseudo-instructions whose biological interpretation raises questions, to say the least. To overcome this difficulty, one option would be to adopt an artificial chemistry model directly inspired by the "genotype-to-phenotype map" and thus allowing to observe experimentally (although *in silico*) the evolution of realistic sequences. The central idea here is that the effects of mutations are, in first approximation, dictated by the structure of genomes and therefore by their decoding process.

I will present Aevol, a simulator based on this idea, as well as some recent results obtained with this model. In particular, I will show how this model can simulate the evolution of populations of genomic sequences along more or less complex phylogenetic trees and how, within the framework of the Evoluthon project, these trees could then be reconstructed using classical (and independent of the model) bioinformatics tools. Without formally constituting a validation of the model, this successful coupling between simulation and phylogenomic analysis suggests that, in Aevol, the sequences have evolved by a process presenting a certain – but unquantifiable – degree of realism.

William Boulton (School of Computing Sciences, University of East Anglia, Norwich, United Kingdom)

Simulating SARS-CoV-2 Sequences: phylogenies and metagenomes

For bioinformaticians, an inspiring challenge arising from the COVID-19 pandemic has been the incredible wealth of data generated by sequencing SARS-CoV-2 genomes from clinical cases, surely the largest epidemiological dataset of its kind [1,2]. In many countries, these data were supplemented by wastewater surveillance, which acted as a bellwether for clinical case numbers and promises to allow for strain-resolved metagenomics to estimate proportions of variants of concern [3].

Many bioinformaticians, including our group, immediately shifted focus to analysing this deluge of new viral sequences, and several groups developed novel algorithms to analyse SARS-CoV-2 phylogenies and metagenomes, such as UShER, iVar and Freyja [4–6]. However, existing tools were not well-equipped to simulate these data, a necessary step when benchmarking new methods. We developed two tools to alleviate these problems: phastSim [7], a fast genome sequence simulator for pandemic scale phylogenies, and SWAMPy [8], a simulator for SARS-CoV-2 metagenomes generated by amplicon sequencing of wastewater samples. This talk will provide a summary of the methods and novel features of these two simulation tools.

References

1. Chen Z., Azman A.S., Chen X. *et al.* (2022) Global landscape of SARS-CoV-2 genomic surveillance and data sharing. *Nat. Genet.* **54**:499–507. doi:10.1038/s41588-022-01033-y
2. Hodcroft E.B., De Maio N., Lanfear R. *et al.* (2021) Want to track pandemic variants faster? Fix the bioinformatics bottleneck *Nature* **591**:30–33. doi:10.1038/d41586-021-00525-x
3. Baaijens J.A., Zulli A., Ott I.M. *et al.* (2021) Variant abundance estimation for SARS-CoV-2 in wastewater using RNA-Seq quantification. *medRxiv* 2021.08.31.21262938. doi:10.1101/2021.08.31.21262938
4. Turakhia Y., Thornlow B., Hinrichs A.S. *et al.* (2021) Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nat. Genet.* **53**:809–816. doi:10.1038/s41588-021-00862-7
5. Grubaugh N.D., Gangavarapu K., Quick J. *et al.* (2019) An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol.* **20**:8. doi:10.1186/s13059-018-1618-7
6. Karthikeyan S, Levy J.I., De Hoff P. *et al.* (2021) Wastewater sequencing uncovers early, cryptic SARS-CoV-2 variant transmission. *medRxiv* 2021.12.21.21268143. doi:10.1101/2021.12.21.21268143
7. De Maio N., Boulton W., Weilguny L. *et al.* (2022) phastSim: Efficient simulation of sequence evolution for pandemic-scale datasets. *PLoS Comput. Biol.* **18**:e1010056. doi:10.1371/journal.pcbi.1010056
8. https://github.com/goldman-gp-ebi/sars-cov-2-metagenomic-simulator