Université Claude Bernard Lyon 1

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

LIRIS

EuroHack 2018
Lugano – October 2018

GPU-Aevol

Guillaume Beslon – Computational Biology
David P. Parsons – Software engineering
Jonathan Rouzaud-Cornabas – High Performance Computing
Mentors: Vasileios Karakasis (ETH Zurich), Jeffrey Kelling (HZDR, Dresden)

Inria
INVENTEURS DU MONDE NUMÉRIQUE

Extremely High Mutation Rate of a Hammerhead Viroid

Selma Gago,[1] Santiago F. Elena,[1] Ricardo Flores,[1] Rafael Sanjuán[1,2]*

Mutation rates vary by orders of magnitude across species (*1*, *2*), with the highest rates measured so far corresponding to RNA viruses (*3*), but little is known about other RNA replicons. Viroids are plant pathogens with minimal non-protein-coding RNA genomes replicated by host RNA polymerases (*4*). We estimated the mutation rate of *Chrysanthemum chlorotic mottle viroid* (CChMVd), a 399-nucleotide chloroplastic viroid with hammerhead ribozymes. Hammerheads are RNA motifs formed by three double-helix regions flanking a core of 15 highly conserved nucleotides critical for catalytic activity (*5*), which mediate self-cleavage of replicative intermediates and, hence, are essential for viroid replication. Hammerhead viroids show elevated genetic variability (*6*), but this variability results

head mutations sampled in vivo, we recreated the mutations by site-directed mutagenesis and assayed for infectivity. Northern-blot hybridizations indicated that plants inoculated with these mutants

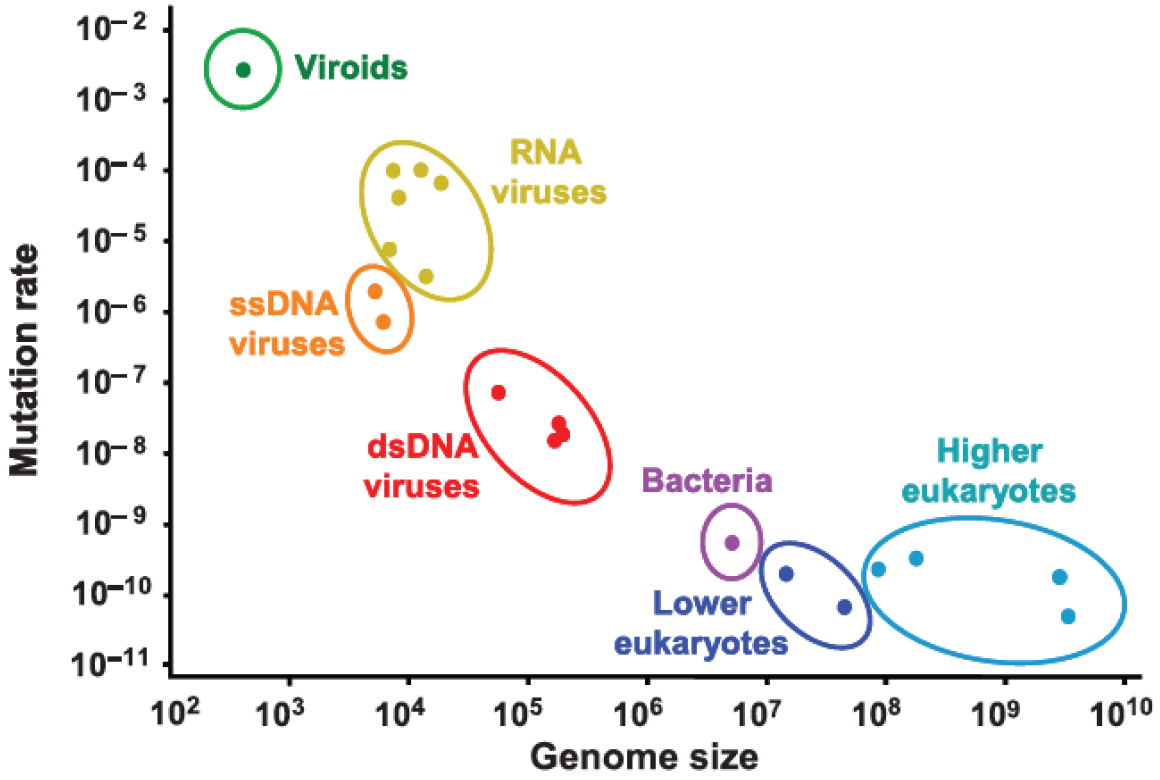

Fig. 1. Per-site mutation rate versus genome size for CChMVd and other biological

(Gago et al., Science, 2009)

Exploring the roots of DNA length and structure by simulation

- Aevol: a simulation software to study the evolution of molecular structure and complexity
  - Individual-Based Model of evolution (~10^2 to 10^4 individuals)
  - Realistic sequence model (10^2 to 10^9 base-pairs)
  - Realistic genotype-to-phenotype map (many motif search on DNA and RNA sequences)
  - Realistic mutation process (local and structural mutations)
  - Long-term evolution experiments (~10^9 generations)
  - Many spontaneous events, rare event fixation (most mutations are deleterious)

→ Typical computation time on a single CPU: 6 months…

→ OpenMP version: 1 month on 32 CPUs

Evident parallelism scheme: distributing individuals… BUT:

- Technical issues:
  - Large variations of organismal complexity → individual-level parallelism is not efficient
  - "Small" population size (~1000 individuals) → smaller than the typical number of processors on a GPU
  - Motif search on long sequences → frequent and demanding memory access
  - Evolution is a historical process → the most efficient parallel scheme may change over time
- Usage constraints:
  - Perfect replicability (to enable rare events post-analysis)
  - Complete history storage (we are interested in the evolutionary process, not in its final result)
  - The selection process requires double-precision computation

What has been done so far…

- Preparing the code for GPU:
  - Changing the Random-Number Generator
  - Packaging all sequence-level operators
  - Optimizing memory usage (pros/cons, CPU/GPU)
  - Basic algorithmic optimizations (pros/cons, CPU/GPU)
  - Code refactoring
- Preparing the environment for efficient testing and validation of the GPU version:
  - Developing tests
  - Code profiling
  - Choice of usecases (demanding ones… but not too much!)
- Debugging Debugging Debugging Debugging Debugging Debugging Debugging Debugging Debugging Debugging Debugging Debugging Debugging Debugging Debugging Debugging

Work in progress…

- First GPU version
  - Dealing with Aevol data-structures…
  - Dealing with Aevol optimizations …
- Finding the best CPU/GPU equilibrium?
  - This may depend on
    - Simulation parameters
    - Macroevolution events (genome growing/streamlining, WGD…)

- Debugging Debugging Debugging Debugging Debugging Debugging Debugging Debugging Debugging Debugging Debugging Debugging Debugging Debugging Debugging Debugging Debugging Debugging Debugging

Final objective

- An efficient GPU version of mini-Aevol (~2,000 code lines) compatible with the full Aevol environment (~70,000 code lines)
  - Must allow post-analyses
  - Must allow GUI integration (possibly offline)
  - Useable by computational biologists
- Bonus: enable exploration of parameter values impossible with the CPU version
  - Typically very large population sizes (e.g. estimated population size of marine Alphaproteobacteria: $2.4\ 10^{28}$ cells)
- Stop Debugging Debugging Debugging Debugging Debugging Debugging Debugging Debugging Debugging Debugging Debugging Debugging Debugging Debugging Debugging