Wright-Fisher Exact Solver GUI (WFES-GUI)

Ivan Krukov, Jason de Koning, Bianca de Sanctis, Alberto Casas-Ortiz

1 Wright-Fisher Exact Solver GUI (WFES-GUI)

Wright-Fisher Exact Solver (WFES) implements a variety of exact calculations with the Wright-Fisher model. Unlike other approaches, WFES does not use simulations or strong simplifying assumptions. WFES benefits from high-performance linear algebra techniques, making it possible to compute exact quantities for biologically realistic population sizes. The following document details the usage of the WFES applications.

Contents

2 Executables

The package consists of several executables, each implementing a different type of a Wright-Fisher model (WF). Each executable has input elements to specify the model parameters and the calculation to be performed.

Most of the executables have additional modes, specifying what type of calculation is to be performed. These modes mostly concern the configuration of absorbing states in the model.

2.1 WFES

The WFES button encapsulates four different executables: WFES Single, WFES Sequential, WFES Sweep and WFES Switching.

2.1.1 `WFES Single`

`WFES Single` implements the standard Wright-Fisher model of a single population. It has the following modes:

- `Absorption` mode assumes that absorption is possible at extinction and fixation boundaries. The calculations assume that the population starts with one or more copies of the allele (see integration for details). This calculates the following statistics:

  Statistics:

  - P_{ext} - probability of extinction.
  - P_{fix} - probability of fixation.
  - T_{abs} - expected number of generations before absorption.
  - $T_{abs\ std}$ - standard deviation of T_{abs}.
  - T_{ext} - expected number of generations before extinction.
  - $T_{ext\ std}$ - standard deviation of T_{ext}.
  - T_{fix} - expected number of generations before fixation.
  - $T_{fix\ std}$ - standard deviation of T_{fix}.

  Matrices and Vectors:

  - Q - transition probability matrix of transient-transient states. This output is not produced by default. Check `Q` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
  - R - transition probability matrix of transient-absorbing states. This output is not produced by default. Check `R` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
  - N - fundamental matrix of the WF model. In this mode, only required rows are calculated. This output is not produced by default. Check `N` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
  - B - matrix result of multiplying N and R. This matrix is useful for obtaining absorption properties of the model. In this mode, only required rows are calculated. This output is not produced by default. Check `B` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
  - N_{ext} - Fundamental matrix of the WF model conditioned by extinction events (Extinction conditional sojourn).
  - N_{fix} - Fundamental matrix of the WF model conditioned by fixation events (Fixation conditional sojourn).
  - I - initial probability distribution of alleles. This output is not produced by default. Check `I` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).

- `Fixation` mode assumes that the extinction boundary is transient, and only the fixation boundary is absorbing. The calculations assume that the population starts with zero copies of the allele. Following statistics are calculated:

  Statistics:

  - $T_{b\ fix}$ - expected number of generations between two fixation events.
  - $T_{b\ fix\ std}$ - standard deviation of $T_{b\ fix}$.
  - $Rate$ - rate of substitutions ($1/T_{b\ fix}$).

  Matrices and Vectors:

- Q - transition probability matrix of transient-transient states. This output is not produced by default. Check `Q` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
- R - transition probability matrix of transient-absorbing states. This output is not produced by default. Check `R` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
- N - fundamental matrix of the WF model. In this mode, only required rows are calculated. This output is not produced by default. Check `N` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
- B - matrix result of multiplying N and R. This matrix is useful for obtaining absorption properties of the model. In this mode, only required rows are calculated. This output is not produced by default. Check `B` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
- I - initial probability distribution of alleles. This output is not produced by default. Check `I` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).

- `Establishment` mode calculates establishment properties. This mode calculates the following outputs:

  Statistics:

  - F_{est} - frequency of establishment.
  - P_{est} - probability of establishment.
  - $T_{seg\ std}$ - standard deviation of T_{seg}.
  - $T_{seg\ ext\ std}$ - standard deviation of $T_{seg-ext}$.
  - $T_{seg\ fix\ std}$ - standard deviation of $T_{seg-fix}$.
  - T_{est} - expected number of generations before establishment.
  - $T_{est\ std}$ - standard deviation of T_{est}.

  Matrices and Vectors:

  - Q - transition probability matrix of transient-transient states. This output is not produced by default. Check `Q` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
  - R - transition probability matrix of transient-absorbing states. This output is not produced by default. Check `R` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
  - I - initial probability distribution of alleles. This output is not produced by default. Check `I` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).

- `Fundamental` mode calculates the entire fundamental matrix of the Wright-Fisher model. There is no assumption about the starting number of alleles. Note that this mode is slow for large matrices ($N > 1000$).

  Matrices and Vectors:

  - Q - transition probability matrix of transient-transient states. This output is not produced by default. Check `Q` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
  - R - transition probability matrix of transient-absorbing states. This output is not produced by default. Check `R` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
  - N - fundamental matrix of the WF model. In this mode, the full matrix is calculated (this is the reason why this mode is slow). This output is not produced by default. Check `N` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
  - V - the variance of the fundamental matrix. This output is not produced by default. Check `V` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
  - I - initial probability distribution of alleles. This output is not produced by default. Check `I` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).

- `Non Absorbing` mode builds a non absorbing matrix of the WF model. This mode calculates the following outputs:

  Matrices and Vectors:

- Q - transition probability matrix of transient-transient states. This output is not produced by default. Check `Q` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
- I - initial probability distribution of alleles. This output is not produced by default. Check `I` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).

- `Equilibrium` mode calculates the equilibrium distribution of allele frequencies. Both boundaries are non-absorbing (this is required for the existence of the equilibrium distribution.).

  Matrices and Vectors:

  - E - equilibrium distribution of allele frequencies. This output is not produced by default. Check `E` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
  - I - initial probability distribution of alleles. This output is not produced by default. Check `I` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).

- `Allele Age` mode calculates moments of the allele age given a current allele frequency. Both extinction and fixation boundaries are absorbing. The calculations assume that the population starts with one or more copies of the allele (see integration for details).

  Statistics:

  - $E(A)$: - expectation of the allele age.
  - $S(A)$: - standard deviation of the allele age.

  Matrices and Vectors:

  - Q - transition probability matrix of transient-transient states. This output is not produced by default. Check `Q` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
  - R - transition probability matrix of transient-absorbing states. This output is not produced by default. Check `R` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
  - I - initial probability distribution of alleles. This output is not produced by default. Check `I` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).

2.1.2 `WFES Switching`

`WFES Switching` implements a time-heterogeneous extension of the Wright-Fisher model - the Markov-modulated Wright-Fisher model. It is possible to switch between different parameter regimes - for example different population sizes, selection parameters, or mutation rates. We refer to each parameter regime as "component". For example, an absorbing model of oscillating population size ($N_1 = 1000, N_2 = 2000$) has two components (corresponding to each population) and $(2N_1 - 1) + (2N_2 - 1) = 7998$ states. The switching between components is parametrized with the initial probability distribution (p), and the rate of switching from one component to the next (r). The following modes are implemented:

- `Absorption` mode assumes that absorption is possible at extinction and fixation boundaries. The calculations assume that the population starts with one or more copies of the allele (see integration for details). This calculates the following statistics:

  Statistics:

  - P_{ext} - probability of extinction.
  - P_{fix} - probability of fixation.
  - T_{ext} - expected number of generations before extinction.
  - $T_{ext\ std}$ - standard deviation of T_{ext}.
  - T_{fix} - expected number of generations before fixation.
  - $T_{fix\ std}$ - standard deviation of T_{fix}.

  Matrices and Vectors:

  - Q - transition probability matrix of transient-transient states. This output is not produced by default. Check `Q` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
  - R - transition probability matrix of transient-absorbing states. This output is not produced by default. Check `R` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).

- N - fundamental matrix of the WF model. In this mode, only required rows are calculated. This output is not produced by default. Check `N` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
- B - matrix result of multiplying N and R. This matrix is useful for obtaining absorption properties of the model. In this mode, only required rows are calculated. This output is not produced by default. Check `B` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
- N_{ext} - Fundamental matrix of the WF model conditioned by extinction events (Extinction conditional sojourn).
- N_{fix} - Fundamental matrix of the WF model conditioned by fixation events (Fixation conditional sojourn).

- `Fixation` mode assumes that the extinction boundary in non-absorbing. Following statistics are calculated:

  Statistics:

  - $T_{b\ fix}$ - expected number of generations between two fixation events.
  - $Rate$ - rate of substitutions ($1/T_{b\ fix}$).

  Matrices and Vectors:

  - Q - transition probability matrix of transient-transient states. This output is not produced by default. Check `Q` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
  - R - transition probability matrix of transient-absorbing states. This output is not produced by default. Check `R` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
  - N - fundamental matrix of the WF model. In this mode, only required rows are calculated. This output is not produced by default. Check `N` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
  - B - matrix result of multiplying N and R. This matrix is useful for obtaining absorption properties of the model. In this mode, only required rows are calculated. This output is not produced by default. Check `B` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).

2.1.3 `WFES Sweep`

`WFES Sweep` implements a type of a switching model with two parameter regimes. The first model is non-absorbing (both extinction and fixation boundaries are transient), and the second model is fixation-only. This is a model of standing genetic variation with pre-adaptive and adaptive components.

There is currently one mode:

- `Fixation` mode assumes that the extinction boundary in non-absorbing. Following statistics are calculated:

  Statistics:

  - $T_{b\ fix}$ - expected number of generations between two fixation events.
  - $Rate$ - rate of substitutions ($1/T_{b\ fix}$).

  Matrices and Vectors:

  - Q - transition probability matrix of transient-transient states. This output is not produced by default. Check `Q` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
  - R - transition probability matrix of transient-absorbing states. This output is not produced by default. Check `R` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
  - N - fundamental matrix of the WF model. In this mode, only required rows are calculated. This output is not produced by default. Check `N` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
  - B - matrix result of multiplying N and R. This matrix is useful for obtaining absorption properties of the model. In this mode, only required rows are calculated. This output is not produced by default. Check `B` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
  - I - initial probability distribution of alleles. This output is not produced by default. Check `I` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).

2.1.4 `WFES Sequential`

`WFES Sequential` implements some calculations by switching over a set of Wright-Fisher models sequentially. There is only one mode where fixation and extinction are both absorbing.

Statistics:

- P_{ext} - probability of extinction.
- P_{fix} - probability of fixation.
- T_{ext} - Expected number of generations before extinction.
- $T_{ext\ std}$ - standard deviation of T_{ext}.
- T_{fix} - Expected number of generations before fixation.
- $T_{fix\ std}$ - standard deviation of T_{fix}.
- $T_{tmo\ std}$ - standard deviation of T_{tmo}.

Matrices and Vectors:

- Q - transition probability matrix of transient-transient states. This output is not produced by default. Check `Q` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
- R - transition probability matrix of transient-absorbing states. This output is not produced by default. Check `R` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
- N - fundamental matrix of the WF model. In this mode, only required rows are calculated. This output is not produced by default. Check `N` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
- B - matrix result of multiplying N and R. This matrix is useful for obtaining absorption properties of the model. In this mode, only required rows are calculated. This output is not produced by default. Check `B` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
- N_{ext} - Fundamental matrix of the WF model conditioned by extinction events (Extinction conditional sojourn).
- N_{fix} - Fundamental matrix of the WF model conditioned by fixation events (Fixation conditional sojourn).

2.2 `WFAF-D`

`WFAF-D` calculates the expected allele frequency distribution for a given piece-wise demographic history. It uses an equilibrium distribution to initiate the calculation, and then iterates forward in time by fast matrix-vector multiplications. It is also possible to start from a given allele frequency distribution. Details on calculation in allele frequency calculation, and details on usage in `WFAF-D` - usage. The following outputs are calculated:

Matrices and Vectors:

- $Dist$ - allele frequency distribution calculated by WFAF-D. Check `Dist` option to direct output to csv file (default), or visualize as a chart (See chart visualization window).

2.3 `WFAF-S`

`WFAF-S` calculates an approximate allele frequency spectrum by leveraging the Markov-modulated Wright-Fisher model. The demography is described by a non-reversible Markov chain, where the transition probabilities are inversely proportional to the expected times in each epoch. The details of implementation in stochastic approximation, and usage in `WFAF-S` - usage. The following outputs are calculated:

Matrices and Vectors:

- Q - transition probability matrix of transient-transient states. This output is not produced by default. Check `Q` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
- N - fundamental matrix of the WF model. In this mode, only required rows are calculated. This output is not produced by default. Check `N` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).

- B - matrix result of multiplying N and R. This matrix is useful for obtaining absorption properties of the model. In this mode, only required rows are calculated. This output is not produced by default. Check `B` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
- $Dist$ - allele frequency distribution calculated by WFAF-S. Check `Dist` option to direct output to csv file (default), or visualize as a chart (See chart visualization window).

2.4 `Time Dist.`

`Time Dist.` iteratively calculates the distribution of time to fixation or extinction. The details on implementation in distribution calculations and usage in `Time Dist.` - usage. The following outputs are calculated:

Matrices and Vectors:

- Q - transition probability matrix of transient-transient states. This output is not produced by default. Check `Q` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
- R - transition probability matrix of transient-absorbing states. This output is not produced by default. Check `R` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
- P - distribution of time to fixation and extinction. It outputs a CSV containing probability distributions over time. Check `P` option to direct output to csv file (default), or visualize as a chart (See chart visualization window). The csv file is organized as follows:
  - $First\ column$ - instant of time (generations).
  - $Second\ column$ - probability of extinction in each generation.
  - $Third\ column$ - probability of fixation in each generation.
  - $Fourth\ column$ - probability of absorption (either fixation or extinction) in each generation.
  - $Fifth\ column$ - cumulative probability of absorption.

2.5 `Time Dist. Dual`

`time_dist_dual`

Matrices and Vectors:

- Q - transition probability matrix of transient-transient states. This output is not produced by default. Check `Q` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
- R - transition probability matrix of transient-absorbing states. This output is not produced by default.Check `R` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
- P - distribution of time to fixation and extinction. It outputs a CSV containing probability distributions over time. Check `P` option to direct output to csv file (default), or visualize as a chart (See chart visualization window). The CSV file is organized as follows:
  - $Firstcolumn$ - instant of time (generations).
  - $Secondcolumn$ - probability of extinction in each generation.
  - $Thirdcolumn$ - probability of fixation in each generation.
  - $Fourthcolumn$ - probability of absorption (either fixation or extinction) in each generation.
  - $Fifthcolumn$ - cumulative probability of absorption.

2.6 `Time Dist. Skip`

Analogue of `Phase Type Dist.`, but excluding the mutation time. Implementation details in distribution calculations, usage in `Time Dist. Skip` - usage.

Matrices and Vectors:

- Q - transition probability matrix of transient-transient states. This output is not produced by default. Check `Q` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).

- R - transition probability matrix of transient-absorbing states. This output is not produced by default. Check `R` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
- P - distribution of time to fixation and extinction. It outputs a CSV containing probability distributions over time. . Check `P` option to direct output to csv file (default), or visualize as a chart (See chart visualization window). The CSV file is organized as follows:
  - $Firstcolumn$ - instant of time (generations).
  - $Secondcolumn$ - probability of substitution in each generation.
  - $Thirdcolumn$ - cumulative probability of substitution.

2.7 `Time Dist. SGV`

Distribution of time to substitution, under a model of standing genetic variation. This is a combination of `Phase Type Dist.` calculation with the `WFES Sweep` model. Implementation details in distribution calculations, usage in `Time Dist. SGV` - usage.

Matrices and Vectors:

- Q - transition probability matrix of transient-transient states. This output is not produced by default. Check `Q` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
- R - transition probability matrix of transient-absorbing states. This output is not produced by default. Check `R` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
- P - distribution of time to fixation and extinction. It outputs a CSV containing probability distributions over time. Check `P` option to direct output to csv file (default), or visualize as a chart (See chart visualization window). The CSV file is organized as follows:
  - $Firstcolumn$ - instant of time (generations).
  - $Secondcolumn$ - probability of substitution in each generation.
  - $Thirdcolumn$ - cumulative probability of substitution.

2.8 `Phase Type Dist.`

`Phase Type Dist.` calculates the distribution of time to substitution. It is the fixation-only absorbing analog of `Time Dist.` (the `WFES Single - Fixation` model). Implementation details in distribution calculations, usage in `Phase Type Dist.` - usage. The following outputs are calculated:

Matrices and Vectors:

- Q - transition probability matrix of transient-transient states. This output is not produced by default. Check `Q` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
- R - transition probability matrix of transient-absorbing states. This output is not produced by default. Check `R` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
- P - distribution of time to fixation and extinction. It outputs a CSV containing probability distributions over time. Check `P` option to direct output to csv file (default), or visualize as a chart (See chart visualization window). The CSV file is organized as follows:
  - $Firstcolumn$ - instant of time (generations).
  - $Secondcolumn$ - probability of substitution in each generation.
  - $Thirdcolumn$ - cumulative probability of substitution.

2.9 `Phase Type Moments`

The calculation of moments of the distribution of time to substitution. `Phase Type Moments` implements the calculation described in [Dayar, 2005] (algorithm 1). This calculates an arbitrary number of moments of the phase-type distribution. Usage details in `Phase Type Moments` - usage.

Statistics:

- $Mean$ - mean of the moments.
- Std - standard deviation of the moments.

Matrices and Vectors:

- Q - transition probability matrix of transient-transient states. This output is not produced by default. Check `Q` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
- R - transition probability matrix of transient-absorbing states. This output is not produced by default. Check `R` option to direct output to csv file (default) or visualize as image (See matrix visualization window for details).
- *Moments* - moments of the distribution of time to substitution. Check `R` option to direct output to csv file (default).

3 Computation

This section explains computations performed in `WFES` in detail.

The main feature of `WFES` is to compute rows of the fundamental matrix of the Wright-Fisher model. From the fundamental matrix, many properties of interest can be derived. We first describe the calculation applied in `WFES Single`, for a standard WF model.

3.1 Wright-Fisher model

The Wright-Fisher model describes a single bi-allelic locus in a population of fixed size. We denote a as the ancestral allele, and A as the derived, or focal allele. The organisms are diploid, so the total number of chromosomes in a population size N is $2N$. Given i copies of derived allele A at time t, the probability of having j copies in the next generation is:

$$P_{i,j}(t+1) = \binom{2N}{j} \psi_i^j (1-\psi_i)^{2N-j} \tag{1}$$

Above, ψ_i is the binomial sampling probability for the number of individuals in the next generation. In the simple case of no mutation or selection, ψ_i only depends on the current number of copies, $\psi_i = \frac{1}{2N}$. One way to parametrize the model with mutation and selection is:

$$\psi_i = \frac{[w_{AA}p^2 + w_{Aa}q](1-\mu_{A\to a}) + [w_{Aa}pq + w_{aa}q^2]\mu_{a\to A}}{w_{AA}p^2 + 2w_{Aa}pq + w_{aa}q^2} \tag{2}$$

Above, $w_{..}$ is the selection coefficient for a particular genotype, $\mu_{A\to a}$ is the backward mutation rate, $\mu_{a\to A}$ is the forward mutation rate. Variables p and q are allele frequencies of A and a respectively: $p = i/2N$, $q = 1-p$. The denominator is the average fitness of the population, $\bar{w}$.

Equation 2 can be parametrized in an arbitrary manner. We follow [Kimura, 1964], and assign the following selection coefficients to the genotypes:

Genotype	Fitness
AA	$1+s$
Aa	$1+sh$
aa	1

Above $h \in [0,1]$ is the dominance coefficient. With the above formulation, (2) simplifies to:

$$\psi_i = \frac{[(1+s)p^2 + (1+sh)q](1-\mu_{A\to a}) + [(1+sh)pq + q^2]\mu_{a\to A}}{(1+s)p^2 + 2(1+sh)pq + q^2} \tag{3}$$

3.2 Fundamental matrix calculation

Equation [1] yields a discrete finite-state Markov chain, with time scale in Wright-Fisher generations. State $i = 0$ corresponds to extinction of A, and $i = 2N$ is fixation of A. The model has $2N+1$ states, where $i = 0$ and $i = 2N$ are absorbing, and the rest are transient. The transition probability matrix $\mathbf{P}$ is $(2N+1) \times (2N+1)$. The transition probability matrix can be re-ordered to group the transient-to-transient entries ($\mathbf{Q}$) and transient-to-absorbing ($\mathbf{R}$) entries:

$$\mathbf{P} = \begin{pmatrix} \mathbf{Q} & \mathbf{R} \\ \mathbf{0} & \mathbf{I_2} \end{pmatrix} \tag{4}$$

With two absorbing states, $\mathbf{0}$ is a 2×2 matrix of zeros, $\mathbf{I_2}$ is a 2×2 identity matrix, and $\mathbf{Q}$ is a $(2N-1) \times (2N-1)$ matrix. For any absorbing Markov chain, there exists a fundamental matrix $\mathbf{N}$:

$$\mathbf{N} = \sum_{k=0}^{\infty} \mathbf{Q}^k = (\mathbf{I} - \mathbf{Q})^{-1} \tag{5}$$

Each entry of $\mathbf{N}_{ij}$ is the expected number of generations spent with j copies, given that we started with i copies. Knowing the entries of $\mathbf{N}$ allows to write down many useful absorption properties of

the Markov chain. For example, probability of absorbing in state k, conditional on starting with i copies is found as the $(i,k)^{th}$ entry of:

$$\mathbf{B} = \mathbf{NR} \tag{6}$$

We can use $\mathbf{B}$ to find the expected number of generations in state j, conditional on starting in i and absorbing in k:

$$\mathbf{E}_{i,k}(j) = \frac{\mathbf{B}_{j,k}}{\mathbf{B}_{i,k}}\mathbf{N}_{ij} \tag{7}$$

The conditional time to absorption in state k is then:

$$T_{abs}(k) = \sum_{j=1}^{2N-1} \mathbf{E}_{i,k}(j) \tag{8}$$

And times to extinction and fixation are:

$$\begin{aligned} T_{ext} &= T_{abs}(0) \\ T_{fix} &= T_{abs}(2N) \end{aligned} \tag{9}$$

These are the properties calculated in `WFES Single` in the `Absorption` mode. See WFES Single - usage for more details.

3.2.1 Example

Figure 1 and shows an example of execution of `WFES Single` in the texttttAbsorption mode.

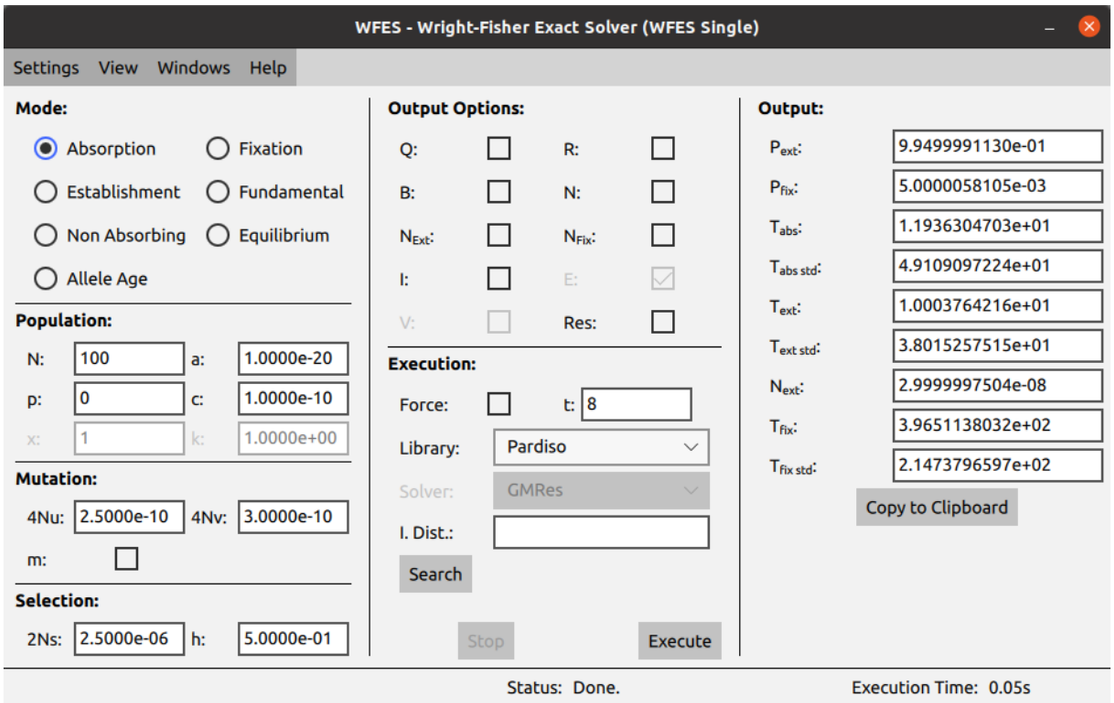

Figure 1: Example of use of `WFES Single` in the `Absorption` mode.

3.3 Solving sparse systems

Solving for the entire matrix $\mathbf{N}$ is expensive for large population size. However, since $\mathbf{N}_{i,j}$ expresses number of generations spent in a state j *conditional* on starting in i, we can simplify the calculation by explicitly conditioning on i. For example if we assume that allele A start with one copy ($i = 1$), then only the first row, $\mathbf{N}_{i,\cdot}$ is of interest. We can generalize this, by assuming a finite forward mutation rate v. In this case, for small $4Nv < 1$, there is a non-zero probability that with $i \leq 1$. However, this probability vanishes quickly with increasing i, and we then only require several rows of $\mathbf{N}$. See more details in integration.

For a starting number of alleles i, we find the i^{th} row of $\mathbf{N}$:

$$(\mathbf{I} - \mathbf{Q})^T \mathbf{N}_{i,\cdot} = \mathbf{I}_{i,\cdot} \tag{10}$$

where $\mathbf{I}_{i,\cdot}$ is the i^{th} column of a $(2N-1) \times (2N-1)$ identity matrix.

This system can be solved by LU decomposition of $(\mathbf{I}-\mathbf{Q})^T$. Once the decomposition is known, we can solve for different right-hand sides of the equation, such as when $i \geq 1$.

To find matrix $\mathbf{B}$, we solve:

$$(\mathbf{I}-\mathbf{Q})\mathbf{B}_{\cdot,0} = \mathbf{I}_{\cdot,0} \tag{11}$$

where $\mathbf{B}_{\cdot,0}$ is the column of $\mathbf{B}$ corresponding to $i = 0$ extinction. Since we have tow absorbing states, we can compute:

$$B_{\cdot,2N} = \mathbf{1} - \mathbf{B}_{\cdot,0} \tag{12}$$

The LU decomposition and solution is performed with `MKL PARDISO` routines. Parameters and settings for the `MKL PARDISO` calls can be found in the source code.

3.4 Entire fundamental matrix

If the entire fundamental matrix is required, it can be calculated with `WFES Single` in the `Fundamental` mode. See `N` option and `V` options in `WFES Single` - usage.

3.4.1 Example

Figures 2 and 3 show an example of calculation and visualization of the fundamental matrix.

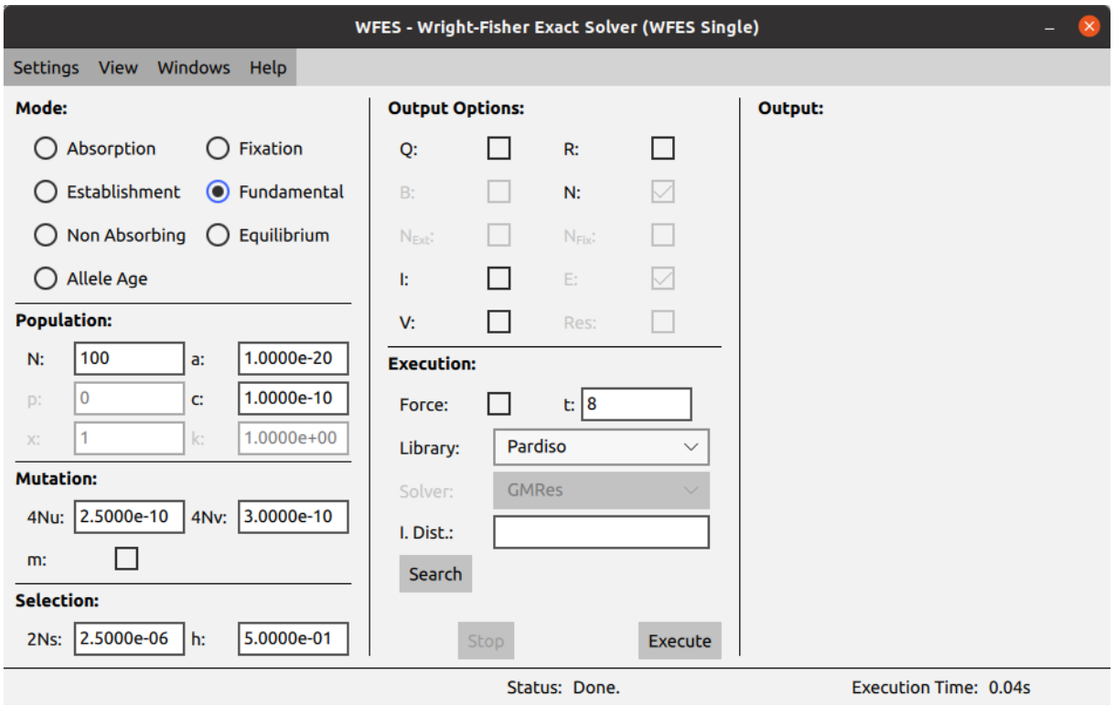

Figure 2: Example of use of `WFES Single` in the `Fundamental` mode.

3.5 Fixation only

The calculation as stated in the previous section applies to the `Absorbing` mode of `WFES Single` - where both extinction and fixation states are absorbing. The other possible mode for the computation is `Fixation` - where only the fixation state ($i = 2N$) is absorbing, and the extinction state ($i = 0$) is transient. In this case, matrix $\mathbf{Q}$ in equation 4 is $2N \times 2N$.

If the extinction state is transient, it can be entered and left many times without terminating the Markov chain. This mode makes it easy to calculate $T_{b\ fix}$ - time between fixations - the total time it takes for a new allele to reach fixation (with the possibility of several extinctions along the way). More details on this calculation an applications can be found in [de Koning and De Sanctis, 2018].

The time between fixations, $T_{b\ fix}$ is calculated in a similar way as T_{fix} for the model with two absorbing states (eq. 6, 7, 8). However, since there is only one absorbing state, no re-conditioning is required (eq. 7). Then the $T_{b\ fix}$ is simply:

$$T_{b\ fix} = \sum_{j=1}^{2N-1} N_{0,j} \tag{13}$$

An advantage of this calculation is that we can safely assume that allele A starts in $i = 0$ copies. Then the integration over the starting number of copies is not necessary, since it is explicitly included in the model as transitions from $i = 0$ to $i > 0$ copies. This then means that we only need to find a single, 0^{th} row of the fundamental matrix.

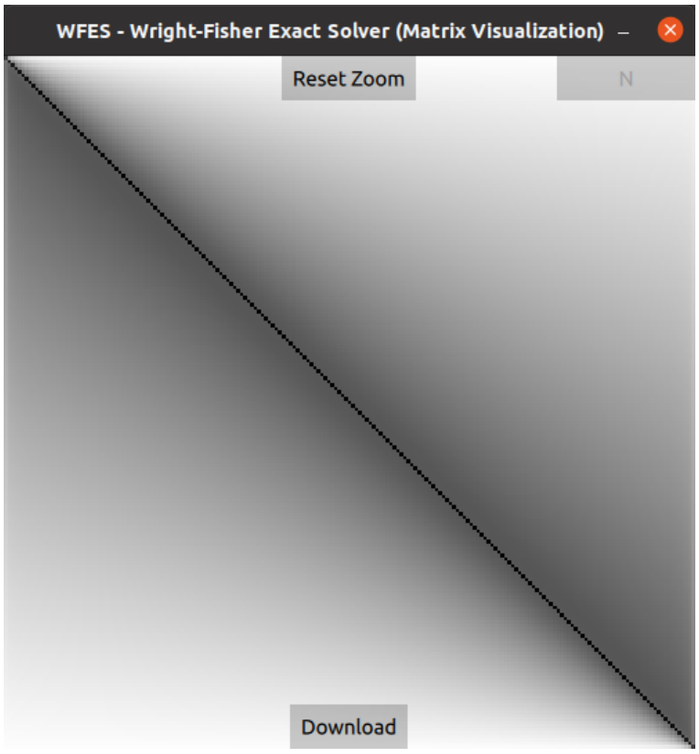

Figure 3: Visualization of fundamental matrix calculated by `WFES Single` in the `Fundamental` mode. The matrix is also generated as a csv file.

3.5.1 Example

Figure 4 shows an example of execution of `WFES Single` in `Fixation` mode.

3.6 Variance calculation

Calculating the variance of the time spent in each state is of interest. It can be found as:

$$\mathbf{N}_{var} = \mathbf{N}(2\mathbf{N}_{dg} - \mathbf{I}) - \mathbf{N}_{sq} \tag{14}$$

where $\mathbf{N}_{dg}$ is the matrix containing the diagonal of $\mathbf{N}$, and $\mathbf{N}_{sq}$ is $\mathbf{N}$ element-wise squared.

If the `V` option is used in the `Fundamental` mode, the entire $\mathbf{N}_{var}$ matrix will be calculated.

In the `Fixation` mode, the standard deviation of $T_{b\ fix}$ is calculated from the first row of $\mathbf{N}$ in equation 13:

$$T_{std} = \sqrt{(2\mathbf{N}_2 - \mathbf{N}_1) - (\mathbf{N}_2)^2} \tag{15}$$

where $\mathbf{N}_1$ and $\mathbf{N}_2$ are found by solving:

$$\begin{aligned} (\mathbf{I} - \mathbf{Q})^T \mathbf{N}_1 &= \mathbf{I}_0 \\ (\mathbf{I} - \mathbf{Q})^T \mathbf{N}_2 &= \mathbf{N}_1 \end{aligned} \tag{16}$$

3.7 Equilibrium allele frequencies

The equilibrium distribution of allele frequencies is one of the key properties of the Wright-Fisher model. We use the method described by [Paige et al., 1975] to solve for the equilibrium distribution (see also [Harrod and Plemmons, 1984]) of a non-absorbing Markov chain. The equilibrium distribution of the Markov chain is defined as vector π, such that $\pi\mathbf{P} = \pi$. This can be expressed in matrix form as:

$$\mathbf{\Pi P} = \mathbf{P} \tag{17}$$

where $\mathbf{\Pi}$ is a $n \times n$ matrix with π in each row. This can be re-written as:

$$\mathbf{\Pi}(\mathbf{P} - \mathbf{I_n}) = \mathbf{0_n} \tag{18}$$

We also have the constraint that $\sum_i \pi_i = 1$, which can be enforced by setting the last columns of $(\mathbf{P} - \mathbf{I_n})$ and $\mathbf{0_n}$, to $e_n = (1, 1, \ldots, 1)^T$. We use the notation $r(A)$ to denote that we set the last column of A to e_n.

$$\begin{aligned} \mathbf{\Pi} r(\mathbf{P} - \mathbf{I_n}) &= r(\mathbf{0_n}) \\ r(\mathbf{P} - \mathbf{I_n})^T \mathbf{\Pi}^T &= r(\mathbf{0_n})^T \end{aligned} \tag{19}$$

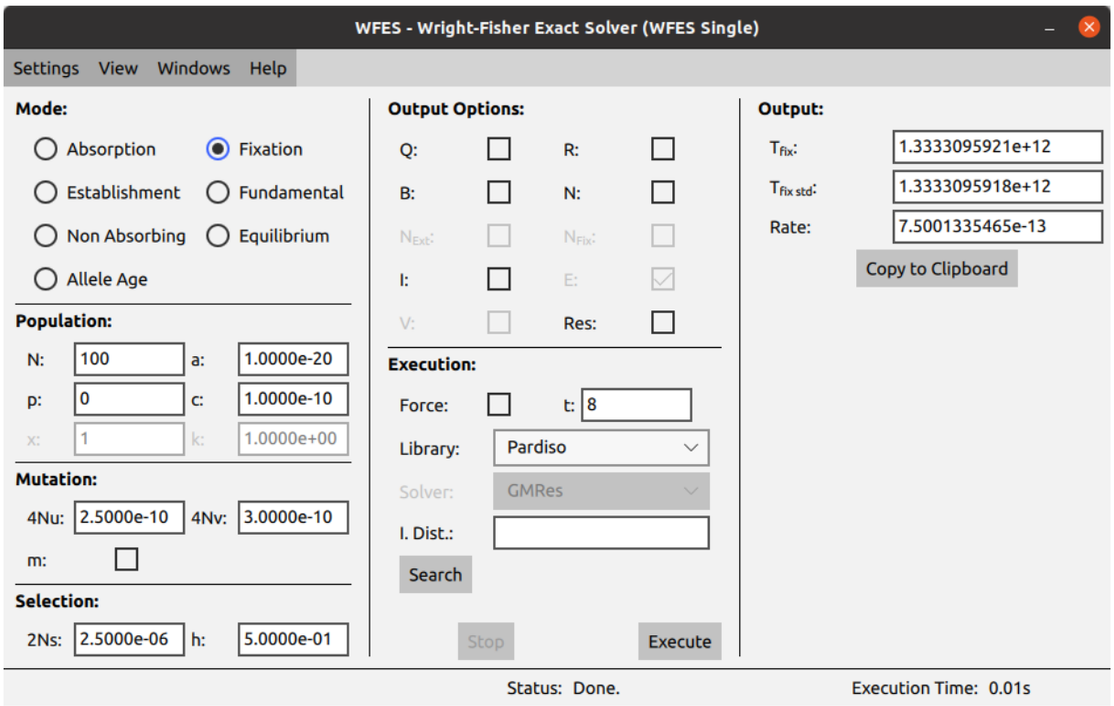

Figure 4: Example of use of `WFES Single` in the `Fixation` mode.

We only require a single row of $\mathbf{\Pi}$. Therefore, we can solve for any row $\Pi_{\cdot,x}$:

$$r(\mathbf{P} - \mathbf{I_n})^T(\mathbf{\Pi}^T)_{\cdot,x} = (r(\mathbf{0_n})^T)_{\cdot,x} \tag{20}$$

This equation is solved with the LU decomposition approach.

Note that the matrix P is a $2N+1$ matrix, since the absorbing states are included. This means that we require that forward and backward mutation rates are non-zero. In case where $\mu_{A\to a} = 0$ or $\mu_{a\to A} = 0$, the matrix P becomes absorbing, and the equilibrium distribution does not exist.

This calculation is performed by `WFES Single` in the `--equilibrium` mode. See `WFES Single` - usage usage for more details.

3.7.1 Example

Figures 5 and 6 show an example of calculation and chart representation of the equilibrium allele frequencies.

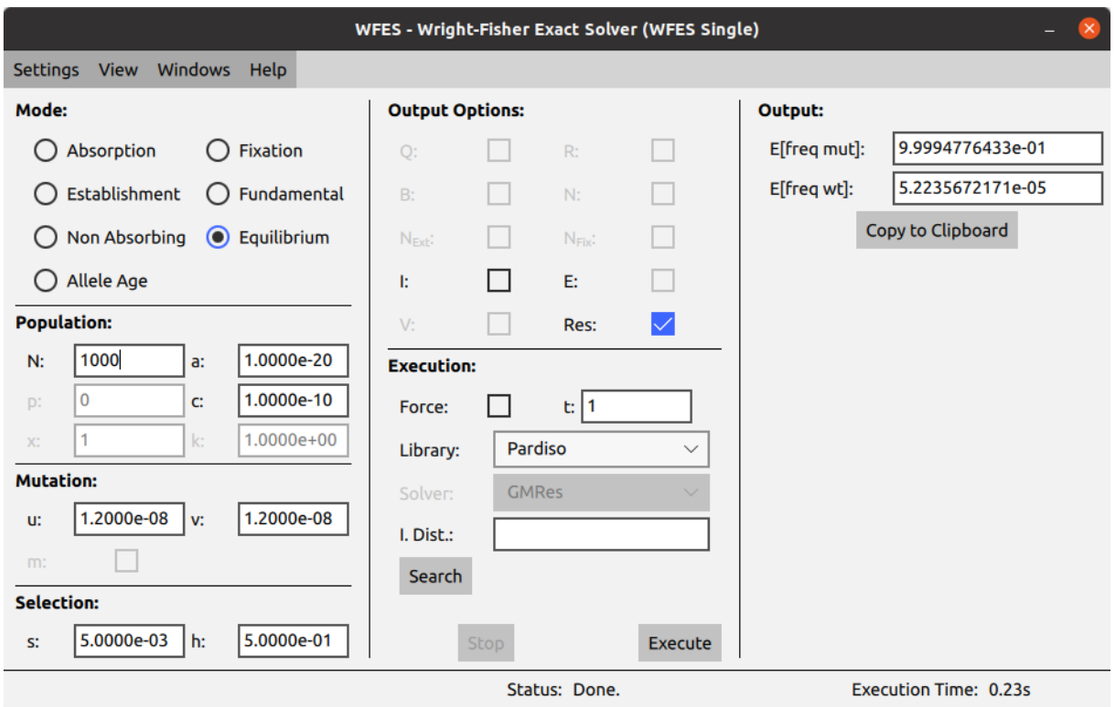

Figure 5: Example of use of `WFES Single` in the `Equilibrium` mode.

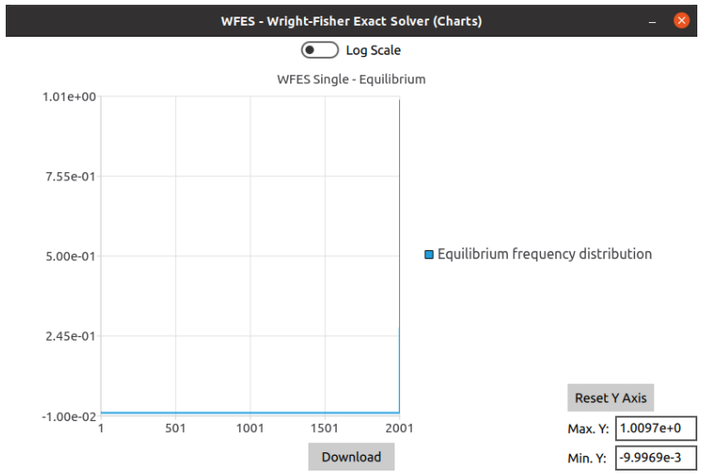

Figure 6: Chart representation of the equilibrium allele frequencies calculated by `WFES Single` in the `Equilibrium` mode. The chart is also generated as a csv file.

3.8 Allele age

For details on the allele age calculation, the user is directed to [De Sanctis et al., 2017]. Briefly, the paper describes a method to find moments of the allele age distribution given an observed number of copies in the WF model. The moments are calculated in an approach similar to those described above.

The calculation is performed by `WFES Single` in the `Allele Age` mode. The observed number of copies is set via the `Observed Copies/x` parameter. See `WFES Single` - usage for more details.

3.8.1 Example

Figure 7 shows an example of calculation of the allele age.

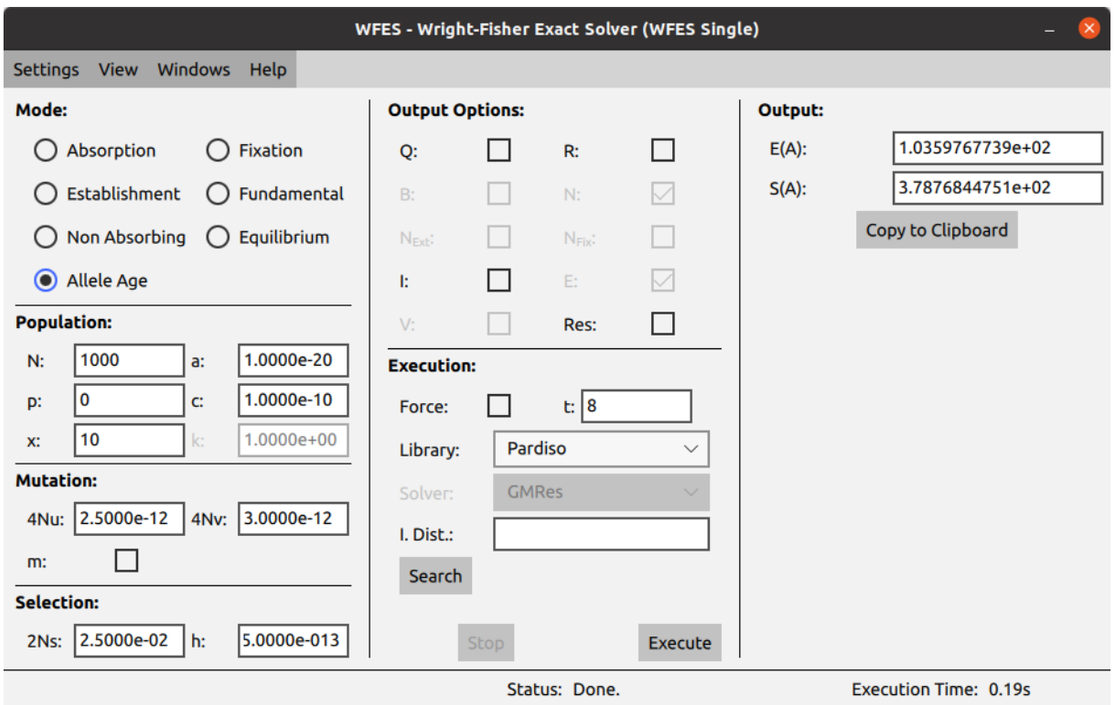

Figure 7: Example of use of `WFES Single` in the `Allele Age` mode.

3.9 Switching models

`WFES Switching` implements an extended time-heterogeneous Wright-Fisher model. The classical WF model describes a single population of constant size. However, this assumption is rarely met in nature. Likewise, classical WF assumes that the rest of the parameters (selection, mutation) are time-invariant. In this section we describe an extension to the Wright-Fisher model with time-variable parameters. We combine a finite set of WF models in a joint Markov-modulated switching process. The switching process assigns a probability of switching between WF component models

with different parameters. Each WF component model can have its own population size, selection coefficient, and mutation rate. Further, `wfes_sweep` combines absorbing and non-absorbing models.

Let $W_1, \ldots, W_n$ represent a finite list of distinct Wright-Fisher components with its own parameter set θ_i. Each component is a full Wright-Fisher Markov model. We also have transition probabilities $r_{x \to y}$ of switching from W_x to W_y at any time. Each component W_i has a transition probability matrix $\mathbf{P}_{(i)}$, which is written in canonical form as (eq. 4):

$$\mathbf{P}_{(i)} = \begin{pmatrix} \mathbf{Q}_{(i)} & \mathbf{R}_{(i)} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \tag{21}$$

We want to describe the join process of switching between $W_1, \ldots, W_n$. We write the canonical form the switching process transition probability matrix as:

$$\mathbf{P} = \left(\begin{array}{cccc|c} \mathbf{Q}_{(1)} & \mathbf{\Gamma}_{(1,2)} & \cdots & \mathbf{\Gamma}_{(1,m)} & \mathbf{R}_{(1)} \\ \mathbf{\Gamma}_{(2,1)} & \mathbf{Q}_{(2)} & \cdots & \mathbf{\Gamma}_{(2,m)} & \mathbf{R}_{(2)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{\Gamma}_{(n,1)} & \mathbf{\Gamma}_{(n,2)} & \cdots & \mathbf{Q}_{(n)} & \mathbf{R}_{(n)} \\ \hline \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{I} \end{array}\right) \tag{22}$$

The $\mathbf{\Gamma}_{x,y}$ matrix defines a matrix of switching from WF component W_x to WF component W_y. The dimensions of the matrix are $(2N_x - 1) \times (2N_y - 1)$ if both W_x and W_y have two absorbing states (or $2N_x \times 2N_y$ if W_x and W_y each have one absorbing state).

The entries of $\mathbf{\Gamma}_{x,y}$ are defined as Wright-Fisher transition probabilities given current i state is in process W_x and next state (j) is in process W_y:

$$\mathbf{\Gamma}_{x,y}(i,j) = \alpha_{x,y} \binom{2N_y}{j} (\psi_{y,i})^j (1 - \psi_{y,i})^{2N_y - j} \tag{23}$$

where $\alpha_{x,y}$ is a switching rate between W_x and W_y.

This formulation essentially matches the frequencies of allele A between different component models. For example if $N_x = 100$ and $N_y = 200$ then $i_x = 10$ would correspond to $i_y = 20$. Note that $\mathbf{\Gamma}_{x,y}(i, \cdot)$ describes a full Wright-Fisher generation, so we are transforming the entire distribution from W_x into W_y.

We additionally use a parameter $p_1, \ldots, p_n$, denoting the probability of starting in each of the components $W_1, \ldots, W_n$. This parameter is most relevant with $\alpha_{x,y}$ imposing a non-reversible switching process.

The calculations on this extended model are similar to those for the simple Wright-Fisher model, since we are still dealing with an finite state absorbing Markov chain. The main difference is that we now deal with extinction and fixation events in each of the component models. To calculate the overall statistic of a model, we weight each component with the probability of starting within each component, p_i. Consider the probability of fixation (probability of extinction is analogous):

$$P_{fix} = \sum_{i=1}^{n} p_i B_{0_i, 2N} \tag{24}$$

where 0_i is the 0^{th} state of the i^{th} component.

These calculations are implemented in `WFES Switching`, with `Absorption` and `Fixation` modes. Matrix parameter α is controlled by `Switching/r` command line flag. See `WFES Switching` - usage for more details.

3.9.1 Example

Figure 8 shows an example of calculation of a reversible model in `WFES Switching`, and Figure 9 shows an example of calculation of a non-reversible model in `WFES Switching`. Both examples use the `Absorption` mode.

3.10 Model of standing genetic variation

`WFES Sweep` implements a model of selection with standing genetic variation. It is a special case of a time-heterogeneous model with two components. The first, pre-adaptive, component is a non-absorbing model with a deleterious or neutral $s_d \leq 0$. This is the model the Markov chain starts in, and the process stays in the first component for an average of τ generations. The process then switches into the second, adaptive, component with $s > 0$. The second component allows fixations with $2N$ copies. Note that the population size in both components is the same.

This model intends to capture the accumulation of standing genetic variation, followed by the onset of positive selection and eventual fixation.

The parameter τ is specified through the rate of transition out of the pre-adaptive component $\lambda = 1/\tau$. The calculations are performed in `WFES Sweep` in `Fixation` mode. See `WFES Sweep` - usage for more detail.

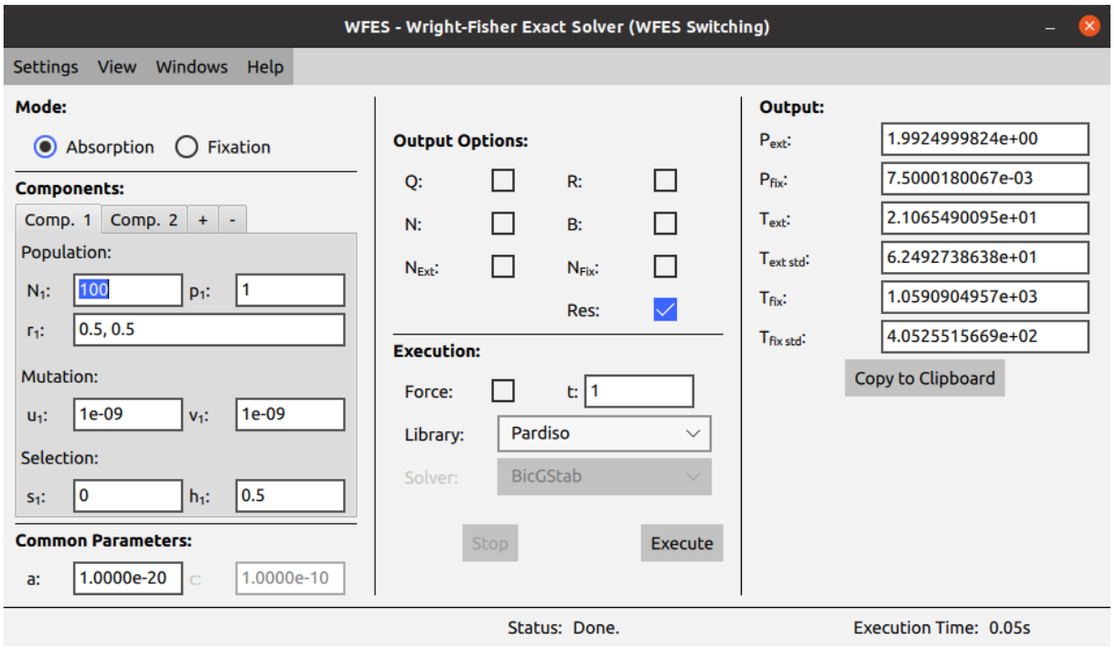

Figure 8: Example of use of `WFES Switching` in the `Absorption` mode. Calculates a reversible model.

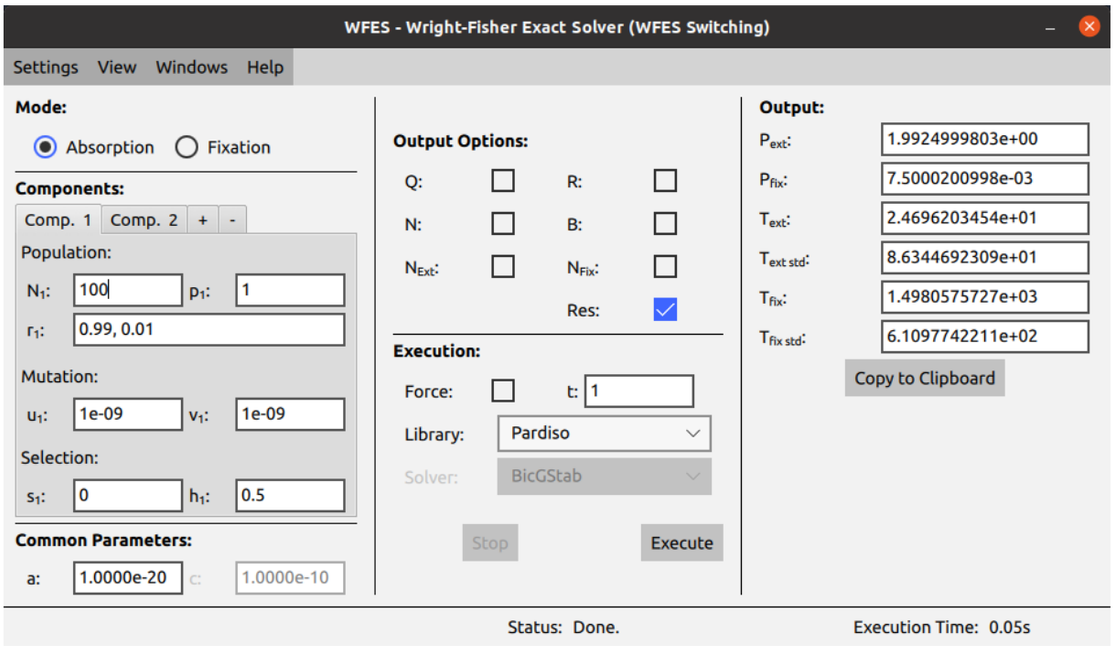

Figure 9: Example of use of `WFES Switching` in the `Absorption` mode. Calculates a non-reversible model.

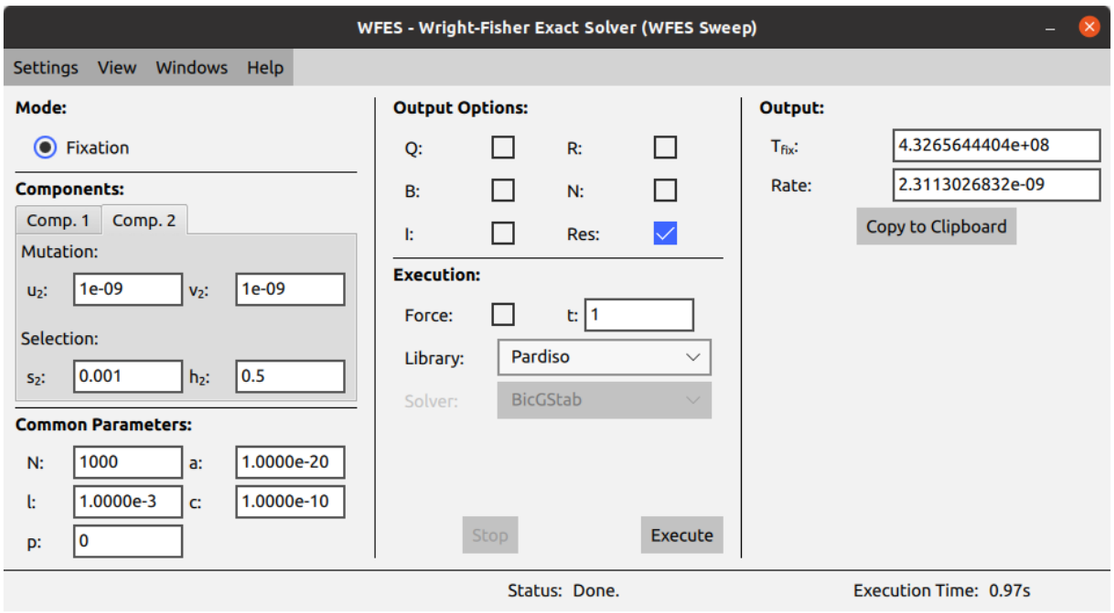

Figure 10: Example of use of `WFES Sweep` in the `Fixation` mode.

3.10.1 Example

Figure 10 shows an example of calculation of `WFES Sweep` in `Fixation` mode.

3.11 Expected allele frequencies with a demographic

`WFAF-D` calculates the expected allele distribution (allele frequency spectrum, AFS) given a piecewise constant demographic history.

The calculation is performed according to the following procedure. Consider a piecewise constant demographic history with population sizes $N_1, N_2, \ldots, N_k$, where each population size epoch lasts for $G_1, G_2, \ldots, G_k$ generations.

1. Acquire the initial probability distribution over allele frequencies for population size N_1. This is done in one of the two ways:
   - Solve for the equilibrium allele frequency distribution using the Paige method, or
   - Read an initial allele frequency distribution from file (specified with `--initial` option).
2. Construct the Wright-Fisher transition probability matrix $\mathbf{Q}_i$ for population size N_i. Multiply the current allele frequency distribution d_i by $\mathbf{Q}_i$ exactly G_i times.
3. Construct a switching transition probability matrix $\mathbf{\Gamma}_{i \to i+1}$. This transition probability matrix incorporates the difference in population size, and other parameters (such as selection). Multiply current d_i by $\mathbf{\Gamma}_{i \to i+1}$
4. Repeat steps 2 and 3 until the final epoch k is reached.

The calculation is feasible since the sparse vector-matrix multiplication in step 2 is relatively cheap.

Detailed usage information is found in `WFAF-D - usage`.

3.11.1 Example

Figure 11 shows an example of calculation of `WFAF-D`.

3.12 Stochastic change in parameters and small population size

`WFAF-S` approximates the calculation of the allele frequency spectrum in `WFAF-D` in two ways. First, the exact epoch transitions are modeled as stochastic transitions, where the expected time of a jump is the same as the length of the epoch. Second, there is an optional approximation of large populations with smaller population models, preserving the population-scaled mutation and selection parameters.

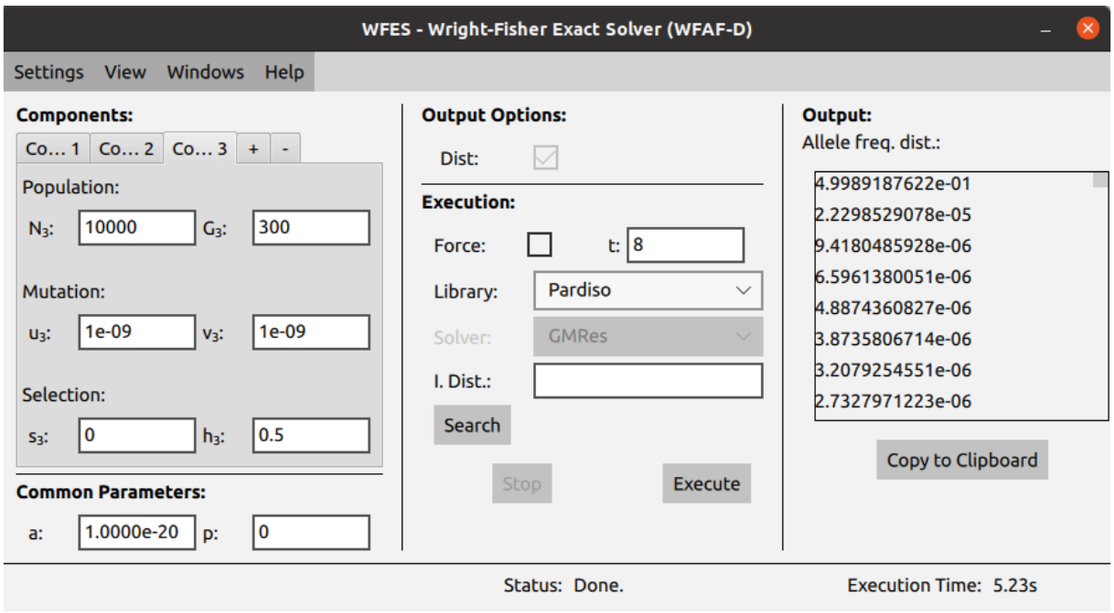

Figure 11: Example of use of `WFAF-D`.

3.12.1 Stochastic epoch transitions

Instead of modeling epoch transitions exactly, we describe them as random variables with the expectation equaling the duration. This allows us to leverage the efficient Markov-modulated Wright-Fisher calculations, which in turn means that that computation time is constant with respect to epoch duration.

The per-generation switching is governed by a Markov chain with transition probability $1/g_i$ from each state to the next, where g_i is the expected number of generations spent in each epoch. Note that each subordinate state is a full Wright-Fisher model with a population size N_i. The probability of staying in the same epoch is then $1 - 1/g_i$, which results in a series of geometrically distributed times. The last state of this Markov chain is absorbing, which is represented by a set of absorbing states, corresponding to an allele count in the final epoch. The allele frequency spectrum is then calculated by finding the absorption probabilities in the final epoch (eq. 11).

3.12.2 Small population size approximation

We can further approximate the allele frequency spectrum by using epochs of smaller population sizes, while preserving the population-scaled values of selection $2Ns$ and mutation $4N\mu$. This corresponds to tracking allele frequencies as opposed to allele counts, as we do in the full model. For example, instead of using a population size of $N = 10,000$, we can use a population of $N_{approx} = 1,000$, with appropriate scaling. This corresponds to an approximating factor of 10.

To improve the behavior of the scaling we can use switching matrix from $N_{approx} = 1,000$, to $N = 10,000$, with dimensions $2N_{approx} \times 2N$, to multiply the AFS we calculate under the approximation. This has the effect of setting the correct support to the resulting distribution.

3.13 Distribution of time to event

`Time *` executables calculate the probability distribution of times to various events in the Wright-Fisher models. While the other programs are concerned with calculation of means or higher moments, here we iteratively calculate the entire distributions.

In general, the probability distribution over the states of a Markov chain with the probability transition matrix $\mathbf{P}$, at time t ($f(t)$), starting at some initial state $f(0)$, is given by:

$$f(t) = f(0)\mathbf{P}^t \tag{25}$$

This can be calculated in an iterative fashion, in order to avoid matrix exponentiation:

$$f(t+1) = f(t)\mathbf{P} \tag{26}$$

Now, consider an absorbing Markov chain, with the transition probability matrix $\mathbf{P}$ partitioned into transient-to-transient probabilities Q and transient-to-absorbing probabilities R. Then, the time-dependent distribution of being in an absorbing state k is given by:

$$P_k(t) = \sum_i \mathbf{R}_{i,k}(f(0)\mathbf{P}^{t-1}) = \mathbf{R}_k \cdot (f(0)\mathbf{P}^{t-1}) \tag{27}$$

Above, $\mathbf{R}_k$ is the k^{th} column of $\mathbf{R}$, which contains the probabilities of absorption from some transient state i into k, within one generation. $(\cdot)$ is the dot product. As before, the calculation can be performed iteratively.

This calculation is the basis of `Time Dist.`, `Time Dist. Skip`, `Time Dist. SGV`, `Time Dist. Dual` and `Phase Type Dist.`.

4 Additional features

4.1 Integration

WFES relies on assumptions about the starting number of copies of an allele in the population. By avoiding the need to calculate the entire fundamental matrix, these assumptions drastically simplify calculations.

4.1.1 Absorbing extinction boundary

Consider a model with two absorbing states - extinction and fixation (`Absorption` mode). The initial configuration can not be 0 copies of allele A, since that is an absorbing state. Thus, the starting number of copies is $i \geq 1$. In the simplest case, we can consider $i = 1$. In this situation, we only require a single row of the fundamental matrix. Alternatively, we can integrate over i by the probability of starting with each number of copies. The conditional probability of starting with i copies of the allele can be derived from the transition probability matrix $\mathbf{P}$:

$$\mathbf{P}_i = \frac{\mathbf{P}_{0,i}}{1 - \mathbf{P}_{0,0}} \quad j \geq 1 \tag{28}$$

The entries of vector $\mathbf{P}_i$ quickly approach zero, and we ignore them below some ϵ. This parameter is set by option `c (Integration Cutoff)`, which is $\epsilon = 10^{-10}$ by default. If option `c (Integration Cutoff)` ≤ 0, no integration is performed, and we assume starting in the smallest starting state ($i = 1$).

We solve equation 10 for any row where $\mathbf{P}_i > \epsilon$, which amounts to several rows with small θ. Since the LU decomposition is the most computationally costly operation, the addition of several rows to the system has minor performance impact.

An alternative approach is to specify the number of copies explicitly. This is done with option `p (Starting copies)`. In this case, only the p^{th} row of the fundamental matrix is found. Note that `p\=1` and `c=-1` are equivalent.

4.1.2 Non-absorbing extinction boundary

In the case where the extinction boundary is not absorbing (`Fixation` mode), the model start with $i = 0$ copies of A. In this case, it is not necessary to integrate over the starting number of copies explicitly - it is automatically included in the model. For the `Fixation` mode, integration flags are ignored.

4.2 Tail truncation

Each row of the WF transition probability matrix is a binomial distribution. To optimize the sparsity of the matrix as a whole, we can consider only the region that contains $1 - \alpha$ mass of the distribution on each row. This truncates the tails of the binomial distribution, significantly increasing the sparsity of the system. The truncation option `a (alpha)` is set to 1×10^{-20} by default. Increasing the value of this parameter will result in faster run times at a sacrifice of precision. In our tests, $\alpha \leq 10^{-15}$ produced results indistinguishable from $\alpha = 0$. With $\alpha = 10^{-5}$, relative error did not exceed 0.03% with $N = 5 \times 10^4$.

4.3 Recurrent mutation

By default, all models in `WFES` allow recurrent mutation during allele segregation. However, this can be turned off with the option `r (no recurrent mutation)`. In this case, the mutation rates u and v describe the rates of only new mutations ($P_{0 \to i}$). No mutations are allowed once there is one or more alleles. Currently, this model is only implemented in `WFES Single`.

4.4 Parameter checks

Before the program executes, the input parameters will be checked for validity. Some of the checks can be skipped by checking the `Force` option. Currently, the following checks are implemented:

- Population size must below 5×10^5 ($N \in [1, 5 \times 10^5]$) - calculations for larger population sizes require excessive amounts of time.
- Selection coefficient must be above -1 ($s \in (-1, 1)$). With the current parameterization (eq. 3), selection coefficients below -1 do not make sense. Positive selection coefficients above 1 can also be problematic, but are currently allowed.

- Mutation rate between 0 and $1/4N$ ($\mu \in (0, 1/4N]$).
  - If the mutation rate is above $1/4N$, then $\theta := 4N\mu > 1$. With higher values of θ, fixation have a conventional meaning. In general, we are calculating statistics concerning first hitting time, which is not the same as fixation.
  - For models where both extinction and fixation boundaries are absorbing, mutation rates can be equal to 0. However, if the extinction boundary is non-absorbing (`Fixation` mode), the forward mutation rate can not be 0. Otherwise, $\mu_{a \to A} = 0$ implies an absorbing extinction boundary, which violates model assumptions. Likewise, if neither of the boundaries are absorbing (`Equilibrium` mode) both forward and mutation rates should be above 0.

5 Graphical User Interface

This version of WFES comes with a Graphical User Interface (GUI) implemented using the Qt Toolkit. In the following sections the different views of the GUI are explained.

5.1 Main Window

When the application is open, the Main Window (Figure 12 appears. This window shows the logo of the application, the Manual Button, the About Button, and a set of buttons corresponding to the different executables.

- The Manual button opens an internet browser and redirects automatically to the Github Wiki of the application. The Wiki is a web version of this manual.
- The About button shows information about the license of WFES, information about the third party libraries, and a link to our laboratory.
- The executable buttons open the windows of the different executables. The WFES button shows the four WFES executables.

Figure 12: Main Window of WFES.

5.2 Executable Window

When an executable is opened, the window corresponding to that executable appears. In Figure 13 it can be seen that an executable window is formed by an upper menu, three columns (two in `Time Dist.`, and a status bar.

- The Upper Menu shows different settings and options. This Menu is common to every executable.
- The first column shows the input options available for the opened executable (see executables and [sec:usage]usage for more information about the options available for each executable).
- The second column shows output options available for the opened executable (see executables for more information about the options available for each executable), and parameters relative to the execution of the executable, like the number of threads (if using openMP), and the library used for solving the linear systems. The Execute button starts an execution and the Stop button stops it.
- The third column shows the different output values generated for the opened executable (see executables for more information about the values generated for each executable). The copy to clipboard button copies the output values in csv format to the clipboard.
- The status bar shows the current status of the execution and the execution time.

If the user puts the mouse cursor over the label of any input/output options, after 0.5 s a tooltip appears with a short description about that option. If the user inserts an invalid value in any field, the field will be highlighted in red.

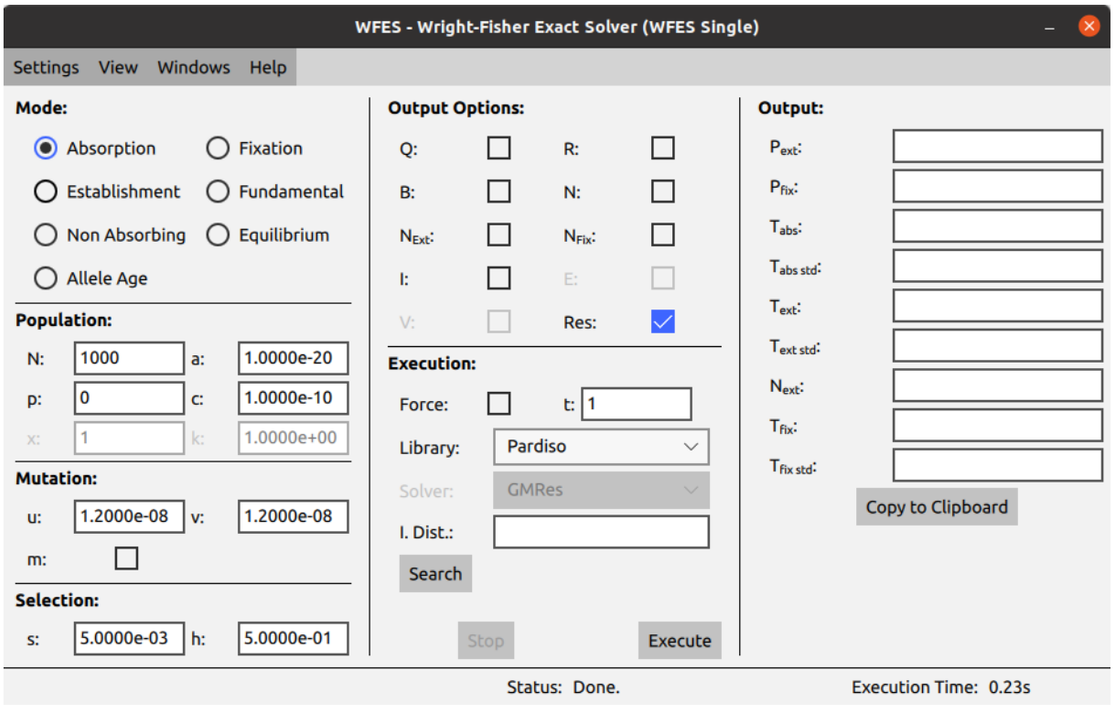

Figure 13: Example of window of an executable. This shows the WFES Single window.

5.3 Upper Menu

The upper menu has some useful options and is common to every executable.

- Settings: Allows the user to save/load a configuration for that executable. Also, the user can activate the use of scaled parameters and the generation of images for the matrices.
- View: Allows the user to open the matrix visualization window and the charts window.
- Window: Allows the user to quickly switch between opened executables windows.
- Help: Allows the user to open the manual of the application and to see information about the license and the development of the application.

5.4 Matrix Visualization Window

Matrices generated by the different executables can be visualized by firs, checking the `Generate Images` option in the upper menu (Settings → Generate Images), checking the requested matrices in the `Output Options` section, and executing the executable. Then, the Matrix Visualizer can be opened by clicking (View → Matrix Visualization). The Matrix Visualizer can be seen in Figure 14.

The generated images can be downloaded by clicking the `Download` button. Also, the user can move the images with the mouse and zoom in/out with the mouse wheel. The zoom and position on the image can be restored by clicking the `Reset Zoom` button.

At the right, the user can click the different buttons that appear to change between the different images generated.

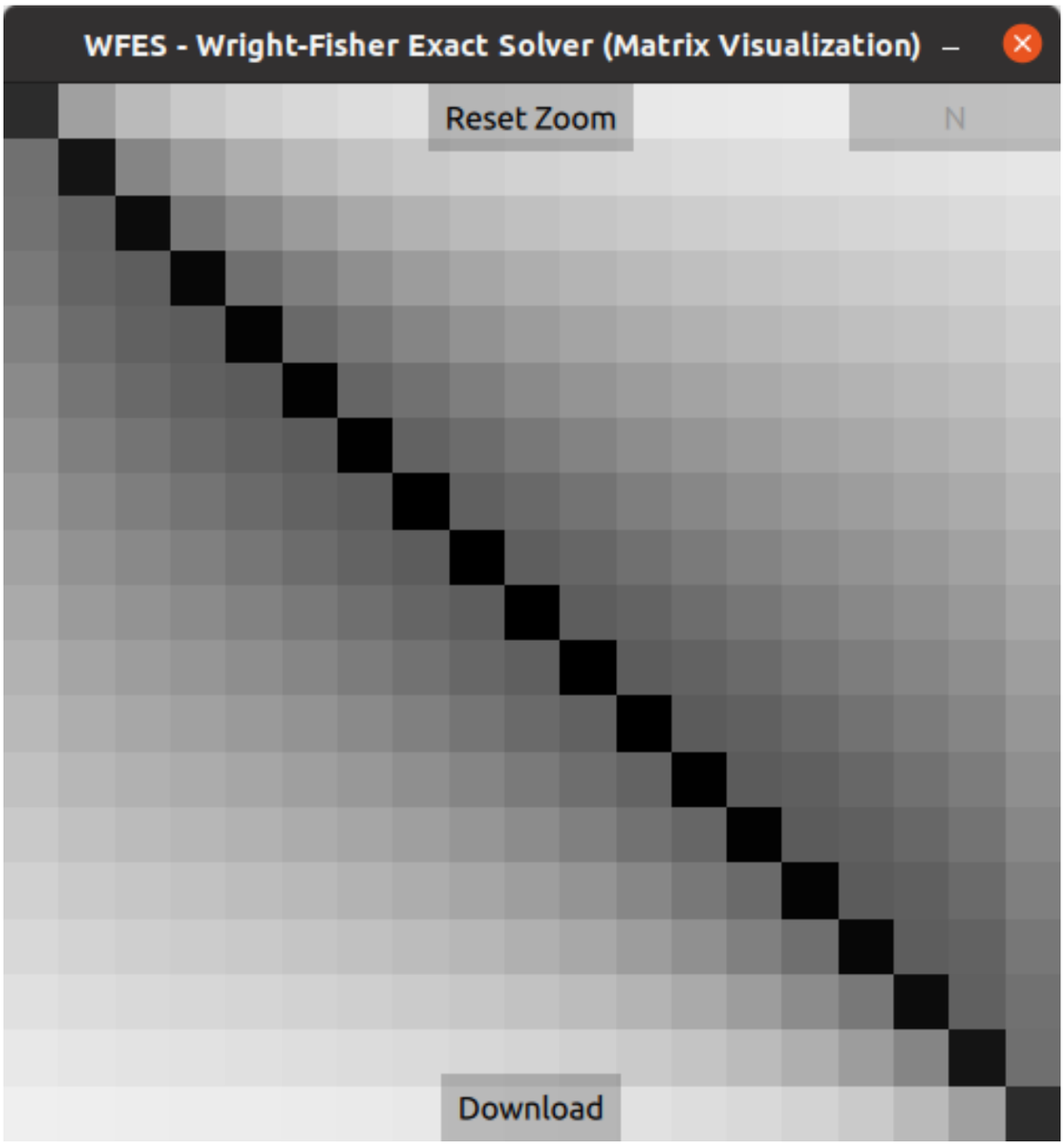

Figure 14: Example of Matrix Visualizer.

5.5 Charts Window

Some executables, like `Time Dist.`, `WFAF-S` and `WFAF-D` are used to generate temporal distributions that can be visualized as charts. The `Charts Window` facilites this.

This window pop ups automatically when one of those executables is executed, but can also be opened from the upper menu (View → Chart Visualization).

As it can be seen in Figure 15, the `Chart Window` has a button at the bottom for Downloading the charts in svg format. At the top of the window, there is a button selecting the scale mode of the x axis of the chart (Linear or logarithmic). The user can set the maximum and minimum values for the Y axis using the input fields at the bottom-right of the window. At the top-right of the window, there are buttons for changing between the different charts generated.

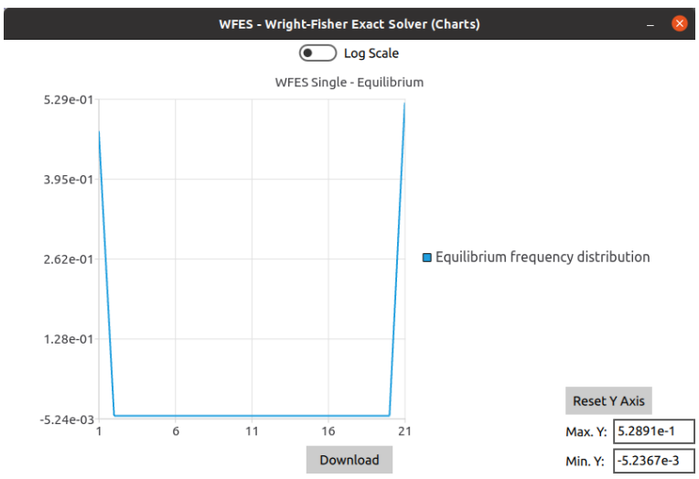

Figure 15: Example of Chart Visualizer.

5.6 Managing configurations

The configurations inserted by the user are automatically saved by the application when the user closes an executable window, and automatically loaded when the application starts. However, the user can save a configuration on the upper menu (Settings → Save Configuration...), and load it (Settings → Load configuration). The configuration files are stored in the Documents folder in Windows, Mac and Linux (/Documents/Wfes), and are human readable. However, we recommend not to modify those files manually.

6 Usage

6.1 `WFES Single` usage

`WGES Single` implements calculations for the standard Wright-Fisher model.

6.1.1 Modes

`WFES Single` supports seven modes:

- `Absorption` mode assumes that absorption is possible at extinction and fixation boundaries.
- `Fixation` mode assumes that the extinction boundary is transient, and the fixation boundary is absorbing.
- `Establishment` mode calculates establishment properties.
- `Fundamental` mode calculates the entire fundamental matrix of the Wright-Fisher model.
- `Non Absorbing` mode builds a non absorbing matrix of the WF model.
- `Equilibrium` mode calculates the equilibrium distribution of allele frequencies.
- `Allele Age` mode calculates moments of the allele age given a current allele frequency.

6.1.2 Options

Table 1: Command line arguments for `WFES Single`

Parameter	Option	Default	Type	Range	Description
Population size	`N`	Required	`int`	$[2, 5 \times 10^5]$	Size of the population
Tail truncation	`a`	`1e-20`	`float`	$[0, 10^{-10}]$	Tail truncation cutoff
Initial number of copies	`p`		`int`	$[1, N]$	Initial number of copies
Integration cutoff	`c`	`1e-10`	`float`	$[0, 10^{-3}]$	Integration cutoff for initial number of copies

Continued on next page

Continued from previous page

Parameter	Option	Default	Type	Range	Description
Observed number of copies	`x`		`int`	$[1, N]$	Observed number of copies for allele age calculation (`Allele Age` only)
Odds ratio	`k`		`float`	$[0, 1]$	Odds ratio (`Establishment` only)
Backward mutation rate	`u`	`1e-9`	`float`	$(0, \frac{1}{4N}]$ *	Backward mutation rate ($A \to a$)
Forward mutation rate	`v`	`1e-9`	`float`	$(0, \frac{1}{4N}]$ *	Forward mutation rate ($a \to A$)
Recurrent mutation	`m`	`true`	`bool`		Exclude recurrent mutation
Selection coefficient	`s`	`0`	`float`	$[-1, 1]$	Individual selection coefficient
Dominance coefficient	`h`	`0.5`	`float`	$[0, 1]$	Dominance coefficient
Initial allele frequency distribution	`i`	Equilibrium	`float[2N+1]`	$[0, 10^{-3}]$	Allele frequency distribution at the start of epoch 1
Number of threads	`t`	`n_cores`	`int`		Number of cores to be used for matrix construction and linear algebra
Q matrix	`Q`		`path`	{`file`, `stdout`}	Output the transition probability matrix for transient states
R matrix	`--output-R`		`path`	{`file`, `stdout`}	Output the transition probability matrix between transient and absorbing states
B matrix	`B`		`path`	{`file`, `stdout`}	Output the conditional absorption probability matrix
N matrix	`N`		`path`	{`file`, `stdout`}	Output the calculated rows of the fundamental matrix
$\mathbf{N}_{ext}$ vector	$\mathbf{N}_{ext}$		`path`	{`file`, `stdout`}	Output extinction-conditional sojourn to file

Continued on next page

Continued from previous page

Parameter	Option	Default	Type	Range	Description
$\mathbf{N}_{fix}$ vector	$\mathbf{N}_{fix}$		`path`	{`file`, `stdout`}	Output fixation-conditional sojourn to file
I matrix	`I`		`path`	{`file`, `stdout`}	Output initial probability distribution
E matrix	`E`		`path`	{`file`, `stdout`}	Output equilibrium distribution (`Equilibrium` only)
V matrix	`V`		`path`	{`file`, `stdout`}	Output variance fundamental matrix (slow)
Force parameters	`Force`		`bool`		Do not perform parameter validity checks
Library	`Library`		`multiple`		Select library for solving matrices.
Solver	`Solver`		`multiple`		Select ViennaCL Solver.

6.2 `WFES Switching` usage

`WFES Switching` implements time-heterogeneous extension to the Wright-Fisher model.

6.2.1 Modes

`WFES Switching` supports two modes:

- `Absorption` - both extinction and fixation boundaries are absorbing for all component models.
- `Fixation` - only fixation boundary is absorbing for all component models.

6.2.2 Options

Table 2: Command line arguments for `WFES Switching`

Parameter	Option	Default	Type	Range	Description
Population sizes	`N`	Required	`int[k]`	$[2, 5 \times 10^5]$	Size of each of the populations
Probability of starting	`p`	`[1/k]*k`	`float[k]`	$[0, 1]$	Probability of starting in each of the component models
Relative probability of switching	`r`	`[1]*[k,k]`	`float[k][k]`	$[0, 1]$	Transition probability matrix between the WF component models
Backward mutation rate	`u`	`[1e-9]*k`	`float[k]`	$(0, \frac{1}{4N}]$ *	Backward mutation rate ($A \to a$)
Forward mutation rate	`v`	`[1e-9]*k`	`float[k]`	$(0, \frac{1}{4N}]$ *	Forward mutation rate ($a \to A$)

Continued on next page

Continued from previous page

Parameter	Option	Default	Type	Range	Description
Selection coefficients	`s`	`[0]*k`	`float[k]`	$[-1, 1]$	Individual selection coefficient
Dominance coefficient	`h`	`[0.5]*k`	`float[k]`	$[0, 1]$	Dominance coefficient
Tail truncation	`a`	`1e-20`	`float`	$[0, 10^{-10}]$	Tail truncation cutoff
Integration cutoff	`c`	`1e-10`	`float`	$[0, 10^{-3}]$	Integration cutoff for initial number of copies (`Fixation` only)
Number of threads	`t`	`n_cores`	`int`		Number of cores to be used for matrix construction and linear algebra
Q matrix	`Q`		`path`	{`file`, `stdout`}	Output the transition probability matrix for transient states
R matrix	`R`		`path`	{`file`, `stdout`}	Output the transition probability matrix between transient and absorbing states
B matrix	`B`		`path`	{`file`, `stdout`}	Output the conditional absorption probability matrix
N matrix	`N`		`path`	{`file`, `stdout`}	Output the calculated rows of the fundamental matrix
$\mathbf{N}_{ext}$ vector	$\mathbf{N}_{ext}$		`path`	{`file`, `stdout`}	Output extinction-conditional sojourn to file
$\mathbf{N}_{fix}$ vector	$\mathbf{N}_{fix}$		`path`	{`file`, `stdout`}	Output fixation-conditional sojourn to file
Force parameters	`Force`		`bool`		Do not perform parameter validity checks
Library	`Library`		`multiple`		Select library for solving matrices.
Solver	`Solver`		`multiple`		Select ViennaCL Solver.

For vector argument defaults, `[z]*k` notation means a vector of length `k`, where each element is `z`. For example, `[z]*3` is `[z,z,z]`.

6.3 `WFES Sweep` usage

`WFES Sweep` implements a model of positive selection with standing genetic variation.

6.3.1 Modes

`WFES Sweep` supports one modes:

- `Fixation` - only fixation boundary is absorbing for the adaptive component

6.3.2 Options

Table 3: Command line arguments for `WFES Sweep`

Parameter	Option	Default	Type	Range	Description
Population sizes	`N`	Required	`int`	$[2, 5 \times 10^5]$	Size of the population
Tail truncation	`a`	`1e-20`	`float`	$[0, 10^{-10}]$	Tail truncation cutoff
Integration cutoff	`c`	`1e-10`	`float`	$[0, 10^{-3}]$	Integration cutoff for initial number of copies
Initial number of copies	`p`		`int`	$[1, N]$	Initial number of copies
Rate of switching	`l`	Required	`float`	$[1e-20, 1]$	Rate of switching from pre-adaptive regime into the adaptive regime
Backward mutation rate	`u`	`[1e-9]*2`	`float[2]`	$(0, \frac{1}{4N}]$ *	Backward mutation rate ($A \to a$)
Forward mutation rate	`v`	`[1e-9]*2`	`float[2]`	$(0, \frac{1}{4N}]$ *	Forward mutation rate ($a \to A$)
Selection coefficients	`s`	Required	`float[2]`	$[-1, 1]$	Individual selection coefficient
Dominance coefficient	`h`	`[0.5]*2`	`float[2]`	$[0, 1]$	Dominance coefficient
Number of threads	`t`	`n_cores`	`int`		Number of cores to be used for matrix construction and linear algebra
Q matrix	`Q`		`path`	{`file`, `stdout`}	Output the transition probability matrix for transient states
R matrix	`R`		`path`	{`file`, `stdout`}	Output the transition probability matrix between transient and absorbing states
B matrix	`B`		`path`	{`file`, `stdout`}	Output the conditional absorption probability matrix
N matrix	`N`		`path`	{`file`, `stdout`}	Output the calculated rows of the fundamental matrix
I matrix	`I`		`path`	{`file`, `stdout`}	Output initial probability distribution
Force parameters	`Force`		`bool`		Do not perform parameter validity checks
Library	`Library`		`multiple`		Select library for solving matrices.
Solver	`Solver`		`multiple`		Select ViennaCL Solver.

For vector argument defaults, `[z]*k` notation means a vector of length `k`, where each element is `z`. For example, `[z]*3` is `[z,z,z]`.

6.4 WFAF-D usage

6.4.1 Input

WFAF-D reads the I Dist. (initial probability distribution) from a file. The file should contain a single-line vector in a .csv format. Spaces around each number are allowed. The number of entries should be $2N + 1$, corresponding to the probability of each allele count for a given population size N.

6.4.2 Output

Unlike other applications, WFAF-D only has one type of output - the allele frequency distribution. The output is generated directly as a csv file. The output is formatted as a single-line csv table.

6.4.3 Options

Table 4: Command line arguments for WFAF-D

Parameter	Option	Default	Type	Range	Description
Population sizes	N	Required	int[k]	$[2, 5 \times 10^5]$	Size of the population
Generations	G	Required	int[k]	$[0, \infty]$	Number of generations each of the k epochs last
Backward mutation rate	u	[1e-9]*k	float[k]	$(0, \frac{1}{4N}]$ *	Backward mutation rate ($A \to a$)
Forward mutation rate	v	[1e-9]*k	float[k]	$(0, \frac{1}{4N}]$ *	Forward mutation rate ($a \to A$)
Selection coefficients	s	[0]*k	float[k]	$[-1, 1]$	Individual selection coefficient
Dominance coefficient	h	[0.5]*k	float[k]	$[0, 1]$	Dominance coefficient
Tail truncation	a	1e-20	float	$[0, 10^{-10}]$	Tail truncation cutoff
Initial allele count	p		int	$[1, 2N]$	Initial number of alleles (I Dist. will be ignored)
Number of threads	t	n_cores	int		Number of cores to be used for matrix construction and linear algebra
Initial allele frequency distribution	I. Dist	Equilibrium	float[2N+1]	$[0, 10^{-3}]$	Allele frequency distribution at the start of epoch 1
Force parameters	Force		bool		Do not perform parameter validity checks
Library	Library		multiple		Select library for solving matrices.
Solver	Solver		multiple		Select ViennaCL Solver.

6.5 WFAF-S usage

WFAF-S approximates the calculation of *AFS* WFAF-D using the Markov-modulated Wright-Fisher model, with optional small population size.

6.5.1 Options

Table 5: Command line arguments for `WFAF-S`

Parameter	Option	Default	Type	Range	Description
Population sizes	`N`	Required	`int[k]`	$[2, 5 \times 10^5]$	Size of each of the populations
Generations	`G`	Required	`int[k]`	$[0, \infty]$	Number of generations each of the k epochs last
Approximating factors	`f`	Required	`float[k]`		Approximating (down) factor for each population size
Backward mutation rate	`u`	`[1e-9]*k`	`float[k]`	$(0, \frac{1}{4N}]$ *	Backward mutation rate ($A \to a$)
Forward mutation rate	`v`	`[1e-9]*k`	`float[k]`	$(0, \frac{1}{4N}]$ *	Forward mutation rate ($a \to A$)
Selection coefficients	`s`	`[0]*k`	`float[k]`	$[-1, 1]$	Individual selection coefficient
Dominance coefficient	`h`	`[0.5]*k`	`float[k]`	$[0, 1]$	Dominance coefficient
Tail truncation	`a`	`1e-20`	`float`	$[0, 10^{-10}]$	Tail truncation cutoff
Initial allele count	`p`		`int`	$[1, 2N]$	Initial number of alleles (`I Dist.` will be ignored)
No projection	`No Proj.`		`bool`		Do not project the distribution down
Number of threads	`t`	`n_cores`	`int`		Number of cores to be used for matrix construction and linear algebra
Initial allele frequency distribution	`I Dist.`	Equilibrium	`float[2N+1]`		Allele frequency distribution at the start of epoch 1
Q matrix	`Q`		`path`	{`file`, `stdout`}	Output the transition probability matrix for transient states
N matrix	`N`		`path`	{`file`, `stdout`}	Output the calculated rows of the fundamental matrix
B matrix	`B`		`path`	{`file`, `stdout`}	Output the conditional absorption probability matrix

Continued on next page

Continued from previous page

Parameter	Option	Default	Type	Range	Description
Force parameters	`Force`		`bool`		Do not perform parameter validity checks
Library	`Library`		`multiple`		Select library for solving matrices.
Solver	`Solver`		`multiple`		Select ViennaCL Solver.

6.6 `Time Dist. *` usage

The following executables have a similar command line interface. They might be merged into a single executable with options in the future.

All the executables require the `P` argument to produce output (activated by default).

6.6.1 `Time Dist.`

This calculates the distribution of time to fixation, extinction.

Output columns

- Time, generations
- Probability of extinction at a given time
- Probability of fixation at a given time
- Probability of absorption (either) at a given time
- Cumulative probability of absorption

6.6.2 `Phase Type Dist.`

This calculates the distribution of time to substitution.

Output columns:

- Time, generations
- Probability of substitution at a given time
- Cumulative probability of substitution

6.6.3 `Time Dist. SGV`

This calculates the distribution of time to substitution under a two-step adaptive model `WFES Sweep`. Note that the parameters are interpreted in the same way (*i.e.* `s` is a vector of size 2).

Output columns:

- Time, generations
- Probability of substitution at a given time
- Cumulative probability of substitution

6.6.4 `Time Dist. Skip`

This calculates the distribution of time to substitution, but ignores the time spent waiting for mutation.

Output columns:

- Time, generations
- Probability of substitution at a given time
- Cumulative probability of substitution

6.6.5 `Time Dist. Dual`

Output columns

- Time, generations
- Probability of extinction at a given time
- Probability of fixation at a given time
- Probability of absorption (either) at a given time
- Cumulative probability of absorption

6.6.6 Options

Table 6: Command line arguments for `Time Dist. *`

Parameter	Option	Default	Type	Range	Description
Population size	`N`	Required	`int`	$[2, 5 \times 10^5]$	Size of the population
Maximum number of generations	`m`	`1e5`	`int`	$[0, 10^{-3}]$	Maximum number of generations to compute the distribution up to
Tail truncation	`a`	`1e-20`	`float`	$[0, 10^{-10}]$	Tail truncation cutoff
Integration cutoff	`c`	`1e-10`	`float`	$[0, 10^{-3}]$	Integration cutoff for initial number of copies
Rate of switching to pre-adaptive phase	`l`	Required	`float`	$[1e-20, 1]$	Rate of switching from pre-adaptive regime into the adaptive regime (`Time Dist. SGV` only)
Recurrent mutation	`r`	`true`	`bool`		Exclude recurrent mutation
Backward mutation rate	`u`	`1e-9`	`float` or `float[2]`	$(0, \frac{1}{4N}]$ *	Backward mutation rate ($A \to a$)
Forward mutation rate	`v`	`1e-9`	`float` or `float[2]`	$(0, \frac{1}{4N}]$ *	Forward mutation rate ($a \to A$)
Selection coefficient	`s`	`0`	`float` or `float[2]`	$[-1, 1]$	Individual selection coefficient
Dominance coefficient	`h`	`0.5`	`float` or `float[2]`	$[0, 1]$	Dominance coefficient
Number of threads	`t`	`n_cores`	`int`		Number of cores to be used for matrix construction and linear algebra

Continued on next page

Continued from previous page

Parameter	Option	Default	Type	Range	Description
Q matrix	`Q`		`path`	{`file`, `stdout`}	Output the transition probability matrix for transient states
R matrix	`R`		`path`	{`file`, `stdout`}	Output the transition probability matrix between transient and absorbing states
P matrix	`P`		`path`	{`file`, `stdout`}	Output phase type distribution
Force parameters	`Force`		`bool`		Do not perform parameter validity checks
Library	`Library`		`multiple`		Select library for solving matrices.
Solver	`Solver`		`multiple`		Select ViennaCL Solver.

6.7 `Phase Type Moments` usage

This calculates arbitrary number of raw moments of the distribution of time to substitution.

6.7.1 Options

Table 7: Command line arguments for `Phase Type Moments`

Parameter	Option	Default	Type	Range	Description
Population size	`N`	Required	`int`	$[2, 5 \times 10^5]$	Size of the population
Number of moments	`k`	20	`int`		Number of moments to output
Tail truncation	`a`	`1e-20`	`float`	$[0, 10^{-10}]$	Tail truncation cutoff
Backward mutation rate	`u`	`1e-9`	`float`	$(0, \frac{1}{4N}]$ *	Backward mutation rate ($A \to a$)
Forward mutation rate	`v`	`1e-9`	`float`	$(0, \frac{1}{4N}]$ *	Forward mutation rate ($a \to A$)
Selection coefficient	`s`	`0`	`float`	$[-1, 1]$	Individual selection coefficient
Dominance coefficient	`h`	`0.5`	`float`	$[0, 1]$	Dominance coefficient
Number of threads	`t`	`n_cores`	`int`		Number of cores to be used for matrix construction and linear algebra
Q matrix	`Q`		`path`	{`file`, `stdout`}	Output the transition probability matrix for transient states

Continued on next page

Continued from previous page

Parameter	Option	Default	Type	Range	Description
R matrix	`R`		`path`	{`file`, `stdout`}	Output the transition probability matrix between transient and absorbing states
Force parameters	`Force`		`bool`		Do not perform parameter validity checks
Library	`Library`		`multiple`		Select library for solving matrices.
Solver	`Solver`		`multiple`		Select ViennaCL Solver.

7 Licenses

This software is licensed under the GNU General Public License Version 3 (GPLv3). The full license text can be found on our GitHub website in the License File and in the installation folder of WFES-GUI. When you use or install this software, you accept the license.

This document, and all of the documents that cannot be cataloged as source code in this project, are licenses under the CC BY-NC-ND License.

This project uses third-party libraries. A list of the third-party libraries used in this project and their licenses can be found in our GitHub website in the Third-Party Licenses section.

References

[Dayar, 2005] Dayar, T. (2005). On moments of discrete phase-type distributions. In *Formal Techniques for Computer Systems and Business Processes*, pages 51–63. Springer.

[de Koning and De Sanctis, 2018] de Koning, A. P. J. and De Sanctis, B. D. (2018). The rate of observable molecular evolution when mutation may not be weak. *bioRxiv*, page 259507.

[De Sanctis et al., 2017] De Sanctis, B., Krukov, I., and de Koning, A. P. J. (2017). Allele age under non-classical assumptions is clarified by an exact computational markov chain approach. *Scientific reports*, 7(1):1–11.

[Harrod and Plemmons, 1984] Harrod, W. and Plemmons, R. (1984). Comparison of some direct methods for computing stationary distributions of markov chains. *SIAM journal on scientific and statistical computing*, 5(2):453–469.

[Kimura, 1964] Kimura, M. (1964). Diffusion models in population genetics. *Journal of Applied Probability*, 1(2):177–232.

[Paige et al., 1975] Paige, C., Styan, G. P., and Wachter, P. G. (1975). Computation of the stationary distribution of a markov chain. *Journal of Statistical Computation and Simulation*, 4(3):173–186.