Large scale simulation of bacterial population evolution over host contact networks

Andreia Sofia Teixeira, Vanda Ribeiro, Pedro T Monteiro,
João A Carriço, Mário Ramirez, Alexandre P Francisco*,

*Corresponding author: aplf@inesc-id.pt

Last modified: June, 2014

Abstract

The understanding of bacterial population genetics and evolution is crucial in epidemic and outbreak studies, but we are not usually able to sample and screen bacterial populations at large and, hence, we must recur to mathematical models, simulations and quantitative analysis. In this report we address the simulation of genetic evolution of bacterial populations in presence of host contact networks. In particular we consider traditional evolution models combined with well mixed and not well mixed host populations, the latter being more realistic. To our knowledge this is the first approach to consider not well mixed host populations, described through complex contact networks. Our results point out that bacterial population diversity can be severely affected by host contact network topology and transmission probabilities alone, without selection phenomena taking place.

1 Introduction

The study of bacterial population genetics is of crucial importance for its understanding in the context of epidemics and outbreaks. The main genetic phenomena that drive the evolution of transmissible pathogens are mutation and recombination. Previous studies have shown that observed population genetic structure of several important human pathogens, such as *Streptococcus pneumoniae* and *Neisseria meningitidis*, can be explained using a simple evolutionary model [1]. This model is based on neutral mutational drift and modulated by recombination, but incorporating the impact of epidemic transmission only for panmictic populations. Although this simple evolutionary model works well for local populations, at a “microepidemic” level, its predictions no longer seem to fit observed genetic relationships of large and widely distributed bacterial populations. With the increasing volume of data obtained with sequence based typing methods, namely by Multi-Locus Sequence Typing (MLST) [6], currently the gold standard for epidemiological surveillance, a much more complex pattern emerges, that cannot be explained solely by the simple ”microepidemic” assumption.

Recently it was shown, using simulated MLST data, that commensal bacteria can be used to infer both local and global properties of the host contact networks of the populations being sampled [9], indicating that the contact network itself can shape the observed pathogen population structures. In particular, authors have shown that, for data simulated from small-world networks, the small world parameter controlling the degree of structure in the contact network can robustly be estimated and that pairwise distances in the network correlate with genetic distances between isolates.

In order to gain new insights into bacterial evolution by simulation on larger scales, we propose in this paper an extension of the above simple evolutionary model, by incorporating the underlying host contact network. Moreover, by implementing an efficient and scalable simulation engine, we are able to simulate large populations over real contact networks. An added value of such model is that it allows the study of the effect of sampling on the perceived bacterial population structure as inferred by traditional methods of phylogenetic inference. The results are providing novel information concerning the impact of host contact network topologies on bacterial population diversity, even in the absence of selection phenomena, that can be used in surveillance of infectious diseases and outbreak investigation and control.

2 The problem

Understanding bacterial population genetics is vital for interpreting the response of bacterial populations to selection pressures such as antibiotic treatment or vaccines targeted at only a subset of strains. The evolution of transmissible bacteria occurs by mutation and localized recombination and is influenced by epidemiological as well as molecular processes [1]. The importance of this fact has become very clear as a fundamental process in strain diversification [10] and as a mechanism by which strains acquire virulence factors or resistance determinants [8].

On the other hand, bacterial population evolution is also influenced by the environment and by host contact networks, through which bacterial populations spread. Note that epidemic spreading is one of most common phenomena in networked systems, e.g. diseases spread from individual to individual through a contact network, and the study of these systems that occur in nature has enabled to obtain reliable data allowing one to quantify the complexity of these networks on which epidemics may propagate.

Hence, which is the impact of host contact network topologies, and associated transmission ratios, on bacterial population evolution and genetic diversity?

3 Approach

While the field of contact network epidemiology is growing fast, it has become critical to develop models and tools to address above problem. In this work we develop a simple and flexible framework that allows the user not only observe the evolution of a MLST population through a configurable network, as also the different behaviours depending on the structure of that network, making it is possible to conclude how mutation and/or recombination affects the population, while given birth to new generations, and the impact on the diversity of each strain. Our main case study will be bacterial population genetics and MLST data, i.e., samples from multilocus sequence typing, that is a technique in which DNA sequences are obtained for seven housekeeping loci and the different sequences at each locus are assigned as different alleles [6]. From an abstract point of view, each pathogen is just characterized through a profile that may be subject to transformations along time, under the influence of genetic events, environment and host contact networks.

Let then each strain in a bacterial population be characterized by MLST, where each sequence type (ST), or MLST profile, is defined by the combination of its seven alleles, a vector of seven integers, where different integers mean different alleles.

Given an underlying host contact network, we simulate a bacterial population at each host, or vertex, with the neutral infinite allele model of Fraser *et al.* [1]. This model is based on a previous null model for evolutionary change, the neutral infinite alleles model (IAM) [4]. We assume non-overlapping generations and, at each step, a new generation is obtained by allowing pathogens to migrate from one vertex to another accordingly to edge transmission probabilities, followed by selection at each host through sampling with replacement from the current host generation. The probability of a ST to occur in the next generation is proportional to its frequency in the current generation after migration. Under IAM, mutation always generates a new allele, leading always to new STs. Recombination, on the other hand, introduces an existing allele randomly selected from the isolates present in the previous generation, which may lead to novel allelic profiles, or the reappearance of existing ones. Mutation or recombination occur independently, with each event being rare and mutation taking precedence over recombination. When a new ST is produced, it is given a new ST number and the parental ST is recorded. For recombination, the allele donor is also recorded.

The model based on the neutral infinite allele model (IAM) just described is also known as the Wright-Fisher model. Neutral evolution means that all individuals have the same fitness. Fitness, in population genetics, is a measure for the expected number of offspring. In the neutral Wright-Fisher model, equal fitness is implemented by equal probabilities for all individuals to be picked as a parent. Each individual will therefore have N chances to become ancestor of the next generations, where N is the size of the population, and in each of these 'trials' the chance that it is picked is $\frac{1}{N}$. That means that the number of offspring of each individual is binomially distributed with parameters $p = \frac{1}{N}$ and $n = N$. In this model, one time step corresponds to a new generation.

Although we do not conducted experiments with alternative models, we have implemented another model, the Moran model. The Moran model describes the evolution of a collection of individuals that is maintained at a constant population size of N. It assumes overlapping generations. At each time, one individual is chosen to reproduce and one individual is chosen to die, so the number of copies of a given allele can go up by one, go down by one, or can stay the same. In the Moran model, it takes N time steps to get through one generation, where N is the effective population size.

4 Methods and results

As described above, our model is based on a contact network where each vertex corresponds to a population of bacteria or to a host, and each edge/link to a connection between populations or hosts. The approach consists in two simulators. Because a population of bacteria evolves over time, even when not in contact with other populations, each node has its own internal simulator – the Single Node Simulator, PopSim – where mutation and recombination rates are used to evolve each population. Regarding interactions between nodes, those are managed by an external simulator - the Network Simulator, NetPopSim - which is responsible for the exchanges of the various populations that are somehow linked through the network.

Since we are using the neutral infinite allele model (IAM), we assume generations are non-overlapping and, in each step, a new generation is obtained by allowing pathogens to migrate from vertex to vertex. This migration is subjected to edge transmission probabilities as we will see below.

Let $G = (V, E)$ be a connected, directed and weighted graph, with $n = |V|$ vertices and $m = |E|$ edges and with an edge transmission probability function $w : E \to I\!R$. Let also g be the number of global generations and δ the step frequency between evolutions and exchanges. The proposed simulator work as follows: The process of exchanging population betweens nodes take three steps:

```
for each generation g do
    run PopSim for each node;
    if $host = false$ && $g = \delta$ then
        run one exchange between nodes;
    else
        run $\delta$ exchanges between nodes;
    end
end
```

1. For each node u create a pool and for each neighbor v of u, gather the individuals from the source node to the pool; the number of individuals to transfer to the pool from each source corresponds to the $w(u, v)$ probability. The pool must also have the individuals corresponding to the minimum self population.

2. After the pool is created, construct a new population with the same size as the previous one, but with the individuals chosen randomly from the pool.

3. After all the new populations have been created, for each node u replace old ones.

Note that the simulator keeps the original populations until the end of the exchanges, replacing them only in the end.

PopSim simulator This simulator is responsible for bacteria evolution on a single host. In the beginning, all the initial bacteria have the same ST signature, that is, all bacteria have the same allele vector. Then PopSim gives birth to new generations through the processes of mutation and recombination. The number of generations created can be defined by the user. PopSim includes several entities:

Profile represents the sequence type of a bacteria. It is a sequence of k alleles where each allele identification number is incremented for each housekeeping gene independently from the other genes of the same type in the same ST. The profile keeps the information about what events originated the new profile, a mutation and/or a recombination.

Population defines a group of bacteria (called individuals) at a given time. The population has a fixed number of individuals throughout the simulation and each individual in the population has a unique identifier. Different individuals can have the same profile even across different generations. This means they belong to the same bacterial strain. The population follows either the Wright-Fisher model or the Moran model.

Individual represents a bacteria in certain a population. Each individual has its own identifier and several individuals can belong to the same bacterial strain, and therefore have the same ST identifier. For each individual the simulator also keeps the identifier for the individual in the previous generation from which it descends (called the parent individual), the ST identifier of the parent individual and the number of recombinations/mutations that originated the current individual.

Allele represents a housekeeping gene. Each existing allele has a unique identifier and has the information of how many recombination and mutation events it has suffered. The allele also keeps the information about its birth generation, that is, the generation in which this allele first appeared. An allele can be present in several generations.

Archive keeps the information about all unique alleles and profiles that exist throughout the simulation. The archive is unique for every simulation, given that it represents all the bacteria known to the simulator.

NetPopSim simulator The external simulator, NetPopSim, is responsible for structuring the network, establish the connections and manage the interactions between nodes. To accomplish this there are two new entities with respect to PopSim:

Node represents a host or a bacterial population. Each node has a maximum number of elements that is defined by the user.

Network represents the network as a directed sparse graph and contains the information about all the links and edge transmission probabilities between nodes.

Once the network is assembled, there are some configurations that need to be made by the user. The NetPopSim is a flexible simulator that allows to test many different configurations. Given the two biological models implemented, you can decide on the probability of the transmission, if the transmission is made with reposition in the origin node, on the number of exchanges before the population evolve, and on several other parameters. Most options are described bellow and can be determined in the properties file of the NetPopSim simulator:

Profile represents the number of alleles of each bacteria.

Generations represent how many times the population of each vertex must evolve, i.e., the number of times that PopSim (the inner simulator) must run.

NodeFile corresponds to the name of the file that contains the network (this must have a specific format detailed later).

Seed represents the seed for the generation of random numbers used in PopSim. This is particular useful if you want to repeat an experience later.

Model represents which biological model to simulate: Wright-Fisher or Moran.

Step represents the number of iterations that one of the simulators runs before the other. Depending on the variable Host it can be or PopSim or NetPopSim.

Host represents a boolean attribute corresponding to the type of the node. If it is a host, then the value is true and NetPopSim runs Step exchanges before running a generation through PopSim, otherwise PopSim runs Step evolutions and, then, exchanges occur through NetPopSim.

Transmission represents a boolean attribute that allows the user to decide if the population evolves with or without reposition. If it is true, then the elements of the population are transmitted without reposition, otherwise they are transmitted but with reposition.

Mutation represents the mutation rate.

Recombination represents the recombination rate.

Test represents the statistical test to run (see details ahead).

Sample represents the frequency with which the data taken from the statistic selected is recorded.

The parameter left to be defined is the transmission probability that is part of the file that contains the network. This file must have the following format:

- The number of nodes and the number of connections in the first line, separated by a tab character.

- A line per node with the node identifier and the population size for each node.

- A line per edge the source identifier, the target identifier, and the transmission probability, separated by tab characters.

After configuring those two files, we are ready to run the simulator. The implementation is in alpha state and it is available upon request. We plan to release it publicly soon.

4.1 Evaluation

To evaluate framework we run it with different types of networks. Some are well known, as the karate and power grid networks available at `http://www-personal.umich.edu/~mejn/netdata/`, some were generated using the tool available at `https://sites.google.com/site/santofortunato/inthepress2`, and others were built following a given topology.

We consider several parameters and the Simpson's index to evaluate diversity. A diversity index is a quantitative measure that quantify the different types, or strains, there are in a dataset. Simpson's Index indicates the probability of two strains sampled randomly from a population belong to two different types, see more details at `http://darwin.phyloviz.net/ComparingPartitions/index.php?link=Tut4`.

In what concerns simulator parameters, we consider different mutation and recombination rates, namely 10/0, 3/10, and 1/10. We consider also different parameterizations for exchange frequency. We considered the basic setting one-to-one where at each evolution it takes place one exchange in the network. Then we tested making 10/100 evolutions and then one exchange, and also 10/100 exchanges and then on evolution.

In terms of size of the population, we fixed the number of elements, independently of the number of nodes. Also, the number of generations (iterations of PopSim) were fixed. To be able to repeat the experiments, we fixed some seeds and also the sampling number to register the evaluation metric. At last, the probability of edge transmission was also fixed. Parameters were as follows:

- **TotalPopulationNetwork** = 50 000
- **Generations PopSim** = 2 500
- **Seed** = 1 / 7 / 11 / 23 / 31
- **Sample** = 10
- **ConnectionProbability** = 0.01

We used several networks to test the framework. We tested with a single node, as the prototype PopSim ran alone itself; we used the Stepping Stone network, where three nodes are connected has a both way path; a triangle network; two cliques linked only by one connection between two nodes of the different cliques, all connections are bidirectional; and the well known karate club network with 34 nodes.

The output consists of several files with the following information: Simpson's index, ST ids, and the Archive that contains all evolution details, including profiles and alleles history.

4.2 Results

Running the network simulator can be a heavy task for a normal computer. While the One Node network can run in a normal computer, using an Intel i7 a 2.3GHz, with 6GB of RAM, running the Karate dataset with some of the possible configurations can turn in an impossible situation. From the experiments made it was possible to see that, even fixing some of the parameters, the time and memory consumed to run an experiment can fluctuate within a wide range. For example, if the user chooses the configuration in which the host parameter is set to 1, this implies that there will be at least the same number of exchanges and of evolutions, if not more. This requires more processing time. Each time that occurs an exchange, you need to create n pools, where n is the number of nodes of the network, sample the pool and then replace the population in each node. It is a process that takes time and if one choose to run 10 or 100 exchanges before an evolution, the time grows almost accordingly. When host is set to 1, the time of execution can vary between 1000 seconds (when exchanges is set to 1 too) and 63000 seconds (when exchanges is set to 100). When running the simulator with host parameter set to 0, it consumes much less time, taking between 300 (exchanges set to 10) and 400 (exchanges set to 100).

In what concerns memory, the consume will depend on recombination and mutation rate. When there is no recombination, only mutation, there are always new STs arising so the archive will grow bigger than in the other cases. Also, there is another point to take into account. At the end of each simulation, the simulator will write the archive all at once in a file, so the memory consumed here, even if only during 5 seconds, is much larger than while running the simulator. From what was observed, in average, the memory consumed may go from 2Gb to 9Gb, but we cannot forget that we are using Java and, hence, memory consumption depends on the garbage collector.

We started to see how mutation/recombination rates could influence the growth of the diversity in only one node. As expected, for higher values of mutation, diversity also increases faster as at each evolution are always generated new STs (Figures 1, 2, and 3).

A simulation using the karate club network (Figure 4) as an example of a weighted graph was run in NetPopSim. Each node had a capacity for 1000 individuals and the simulation was run for 7 loci, similar to a typical MLST profile. The mutation/recombination ratio was set at 3/10 and, at each generation of a Wright Fisher Model, 5% the strains were exchanged through each edge on the graph. The simulation was run for 2500 generation and Simpson's Index of Diversity (SID) was calculated for each node. The run time for this simulation is around 700 seconds. We observed that strain persistence and local evolutionary events reflecting local expansion and limitation of genetic exchange due to the host network, can lead to drastic reductions of SID values that can erroneously be attributed to selection (Figure 5). Distinction between drift and selection is essential if we hope to understand natural evolutionary processes.

When testing networks, if the links are bidirectional and the transmission rate constant, the behaviour is very similar. The only thing that is different is related with the previous observation with the One Node experiment, being the moment when the Simpson's index reaches its maximum. Figures 9 and 10 depict results for the Stepping Stone network and Figures 11 and 12 depict results for the Cliques network. You can observe that the behaviour is very similar. Other result is that if you turn Karate Club network into a bidirectional network (Figures 13 and 14), the diversity have no oscillations contrary to what we observed in Figure 5.

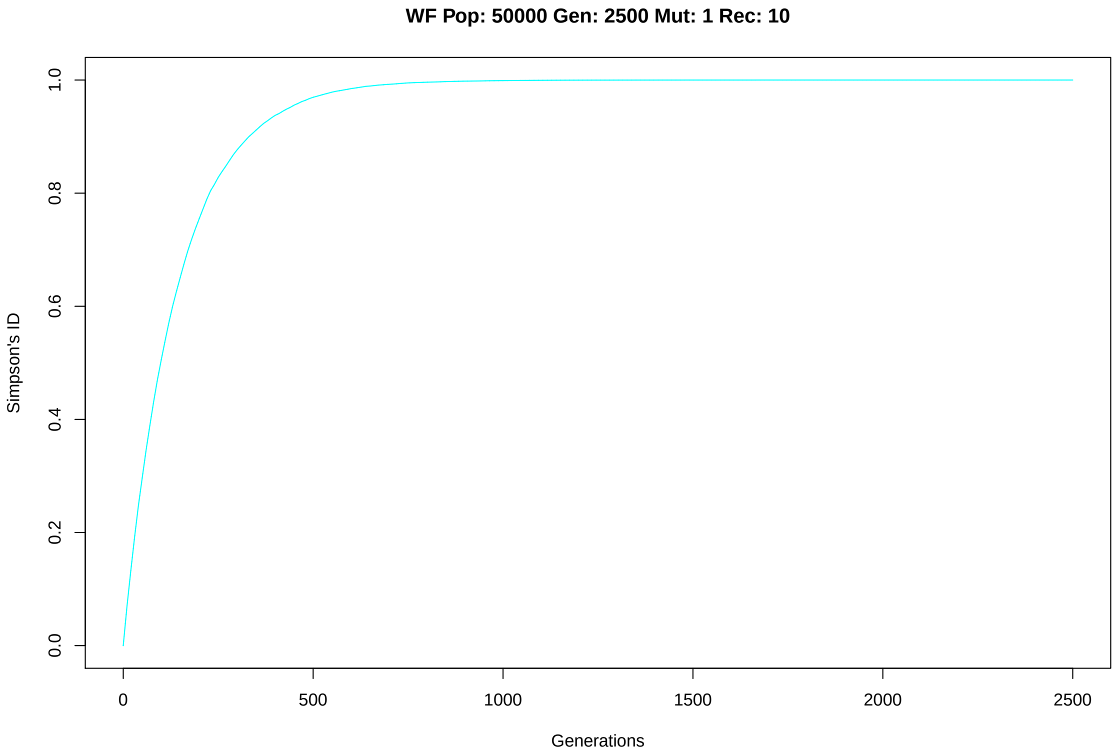

Figure 1: One node, 50000 individuals, 2500 generations, 1/10 mutation/recombination rate.

5 Final Remarks

The dynamic and behavior of complex networks have been studied over the years. Watts in [12] studied the impact of clustering on several processes, including games, cooperation, the Prisoner's Dilemma, cellular automata while Lago-Fernandez in [5] studied synchronization. Wang and Chen in [11] demonstrated that the inhomogeneous scale-free topology plays an important role in determining synchronization in complex networks. Usually these processes can be formally studied with the help of epidemiological models, the so-called susceptible/infective/removed (SIR) models that can be solved exactly on a wide variety of networks [3, 7], but we lack models for a more basic phenomenon, namely for objects evolution and transformation on top of such networks, being them bacterial pathogens, autonomous agents, or just social profiles. In this context, we also lack flexible programming frameworks for the generation and analysis of epidemiological contact networks and for the simulation of disease transmission through such networks.

Recently, Hladish *et al.* developed Epifire [2], an open-source programming interface for the rapid development of SIR network models with a focus in contact network epidemiology, but once again there is a lack of tools to study other kind of models. Epifire has an user friendly interface that allows the user through a point-and-click manner to generate networks, conducting epidemic simulations, and creating figures. This interface is particularly useful as a pedagogical tool.

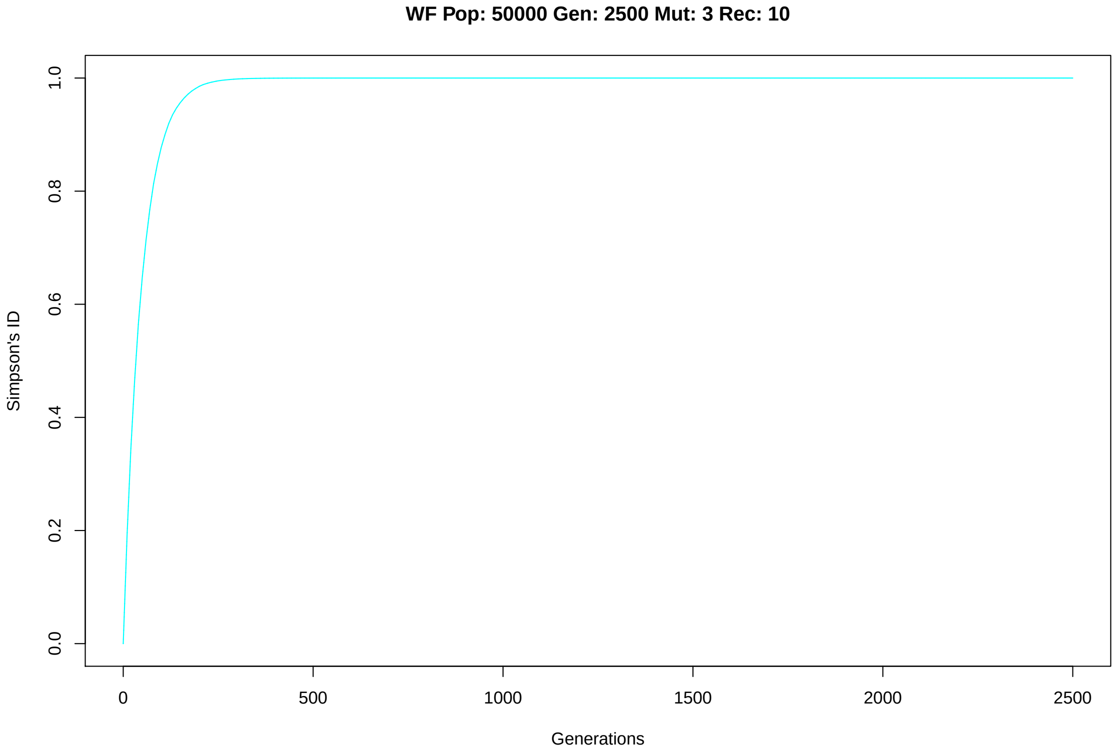

Figure 2: One node, 50000 individuals, 2500 generations, 3/10 mutation/recombination rate.

In this work we proposed and implemented a simulator for population evolution on top of host contact networks, supporting the Wright-Fisher model and the Moran model. Then we conducted several experiments and we were able to observe that strain persistence and local evolutionary events reflecting local expansion and limitation of genetic exchange due to the host network, can lead to drastic reductions of SID values that can erroneously be attributed to selection. This is a first step towards the distinction between drift and selection, which understanding is essential if we hope to understand natural evolutionary processes.

This is an ongoing work and we plan to run our simulators on top of a distributed platform, where we may be able to simulate populations with millions of strains over real-like networks. We are particularly interested on how equilibrium levels of diversity and evolution patterns, as displayed by phylogenetic trees, are affected by small-world effects and community structure within the underlying network. Moreover, we hope to be able to study how failure points and localized sampling procedures allow us to control and detect outbreaks and epidemics.

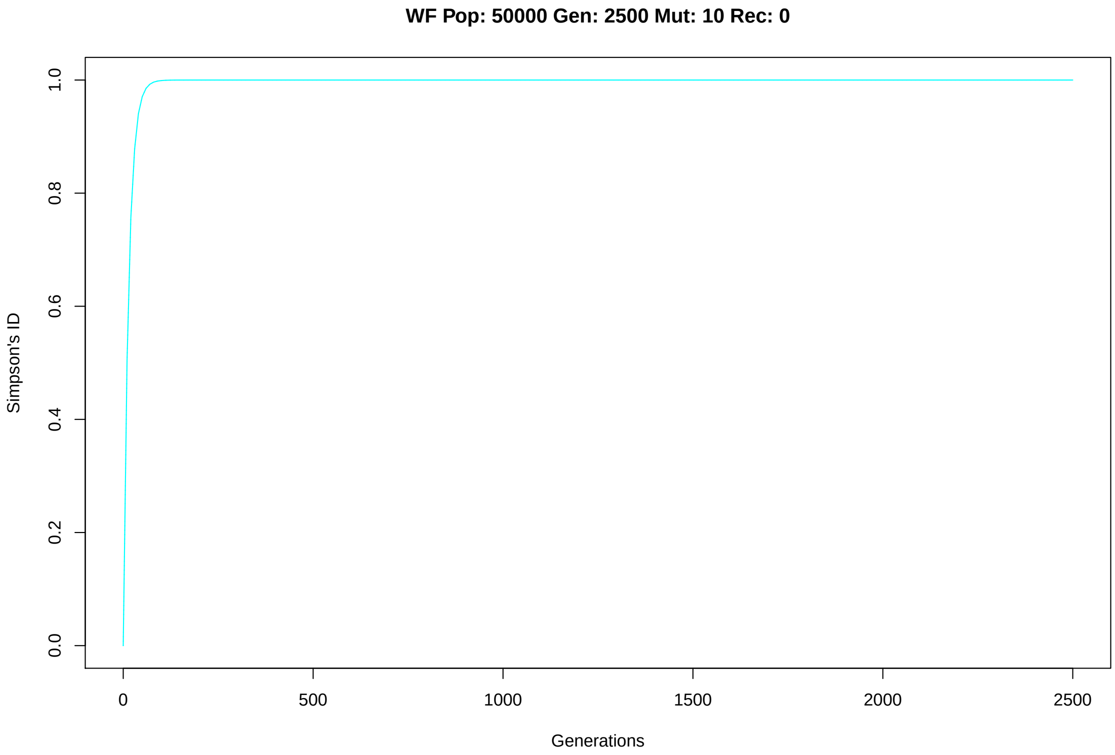

Figure 3: One node, 50000 individuals, 2500 generations, 10/0 mutation/recombination rate.

References

[1] C. Fraser, W.P. Hanage, and B.G. Spratt. Neutral microepidemic evolution of bacterial pathogens. *PNAS*, 102(6):1968–1973, 2005.

[2] J.T. Hladish, E. Melamud, L.A. Barrera, A. Galvani, and L. Ancel Meyers. Epifire: an open source c++ library and application for contact network epidemiology. *BMC Bioinformatics*, 13:76, 2012.

[3] B. Karrer and M. E. J. Newman. A message passing approach for general epidemic models. *Phys. Rev. E*, 82, 2002.

[4] M. Kimura. Evolutionary rate at the molecular level. *Nature*, 217:624–626, 1968.

[5] L. F. Lago-Fernandez, R. Huerta, F. Corbacho, and J. A. Siguenza. Fast response and temporal coherent oscillations in small-world networks. *Phys. Rev. Lett.*, 84(12):2758–2761, 2000.

[6] M.C.J. Maiden, J.A. Bygraves, E. Feil, G. Morelli, J.E. Russell, R. Urwin, Q. Zhang, J. Zhou, K. Zurth, D.A. Caugant, et al. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *PNAS*, 95(6):3140–3145, 1998.

[7] M. E. J. Newman. The spread of epidemic disease on networks. *Phys. Rev. E.*, 66, 2002.

[8] H. Ochman, J. G. Lawrence, and E. A. Groisman. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405:299–304, 2000.

[9] M.M. Pluciński, R. Starfield, and R.P.P. Almeida. Inferring social network structure from bacterial sequence data. *PloS One*, 6(8):e22685, 2011.

[10] B. G. Spratt, W. P. Hanage, and E. J. Feil. The relative contributions of recombination and point mutation to the diversification of bacterial clones. *Curr Opin Microbiol.*, 4(5):602–606, 2001.

[11] X. Wang and G. Chen. Synchronization in scale-free dynamical networks: Robustness and fragility. *eprint arXiv:cond-mat/0105014 of Modern Physics*, 2002.

[12] D. Watts. *Small Worlds: The dynamics of networks between order and randomness.* Princeton, NJ: Princeton University, 1999.

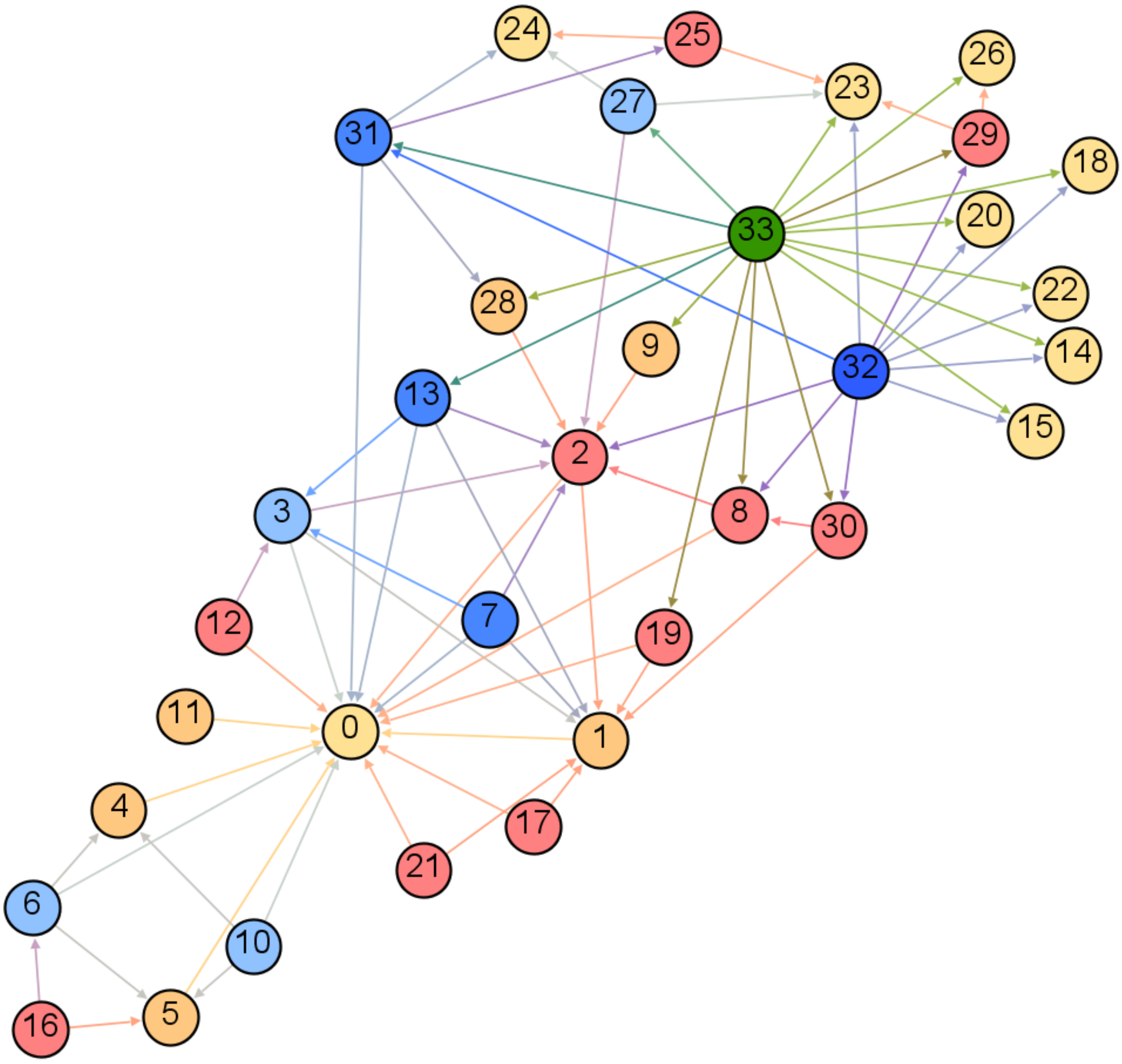

Figure 4: Karate network.

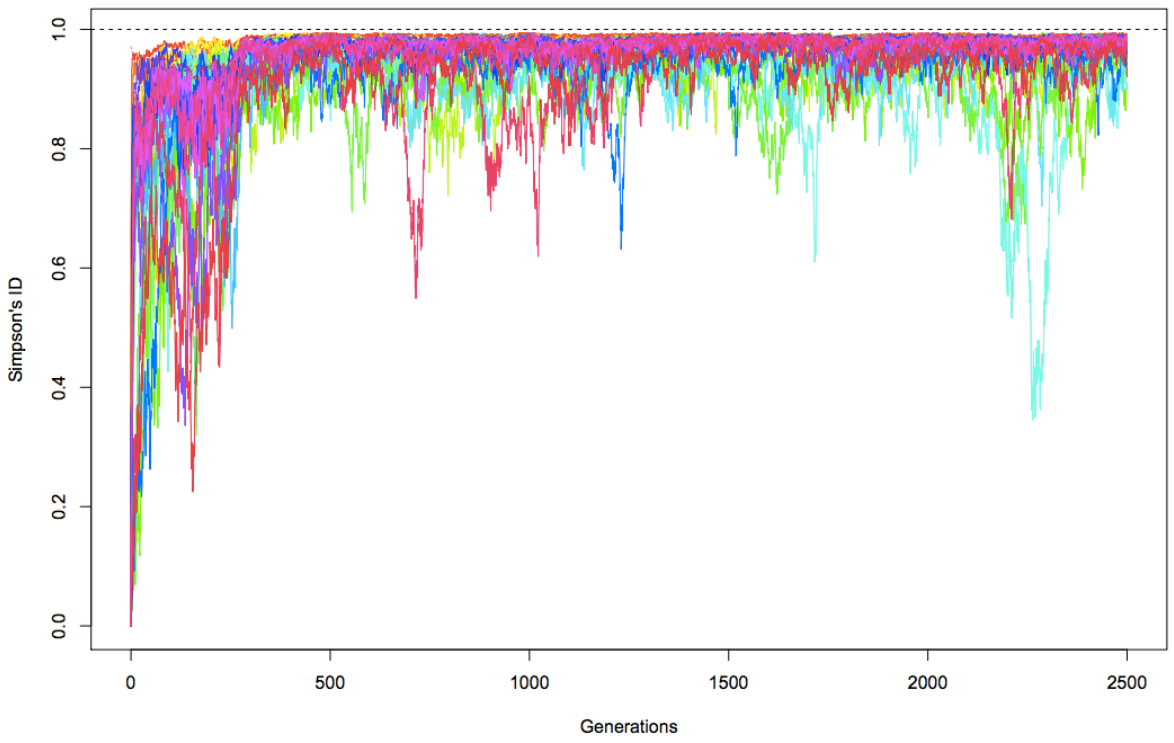

Figure 5: Karate network, diversity for all nodes with a population of 1000 individuals per node evolving for 2500 generations.

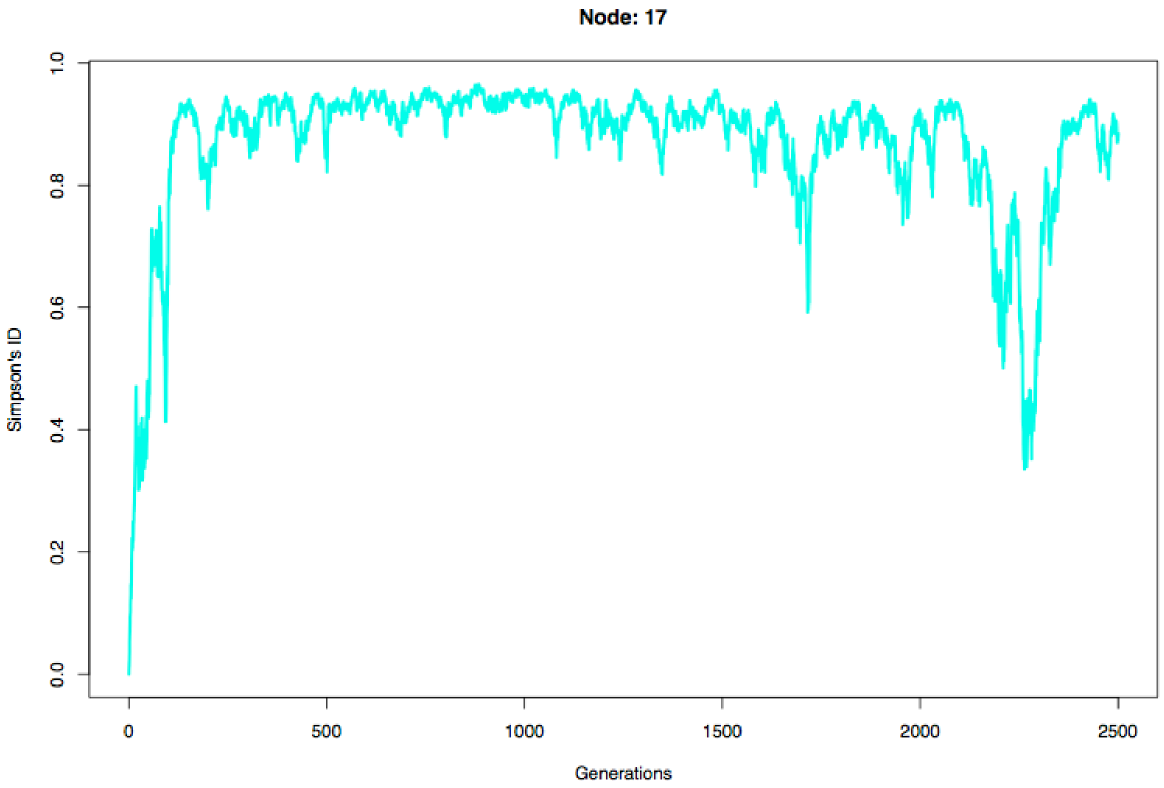

Figure 6: Karate network, diversity for node 17 with a population of 1000 individuals evolving for 2500 generations.

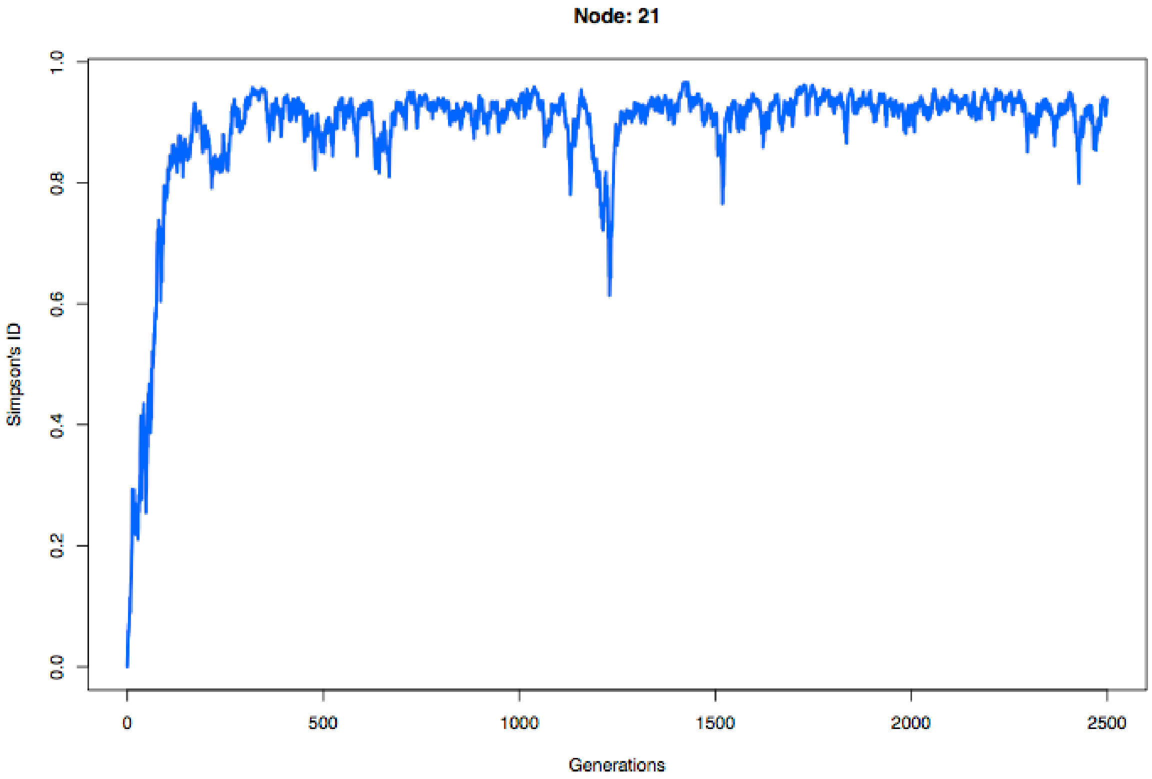

Figure 7: Karate network, diversity for node 21 with a population of 1000 individuals evolving for 2500 generations.

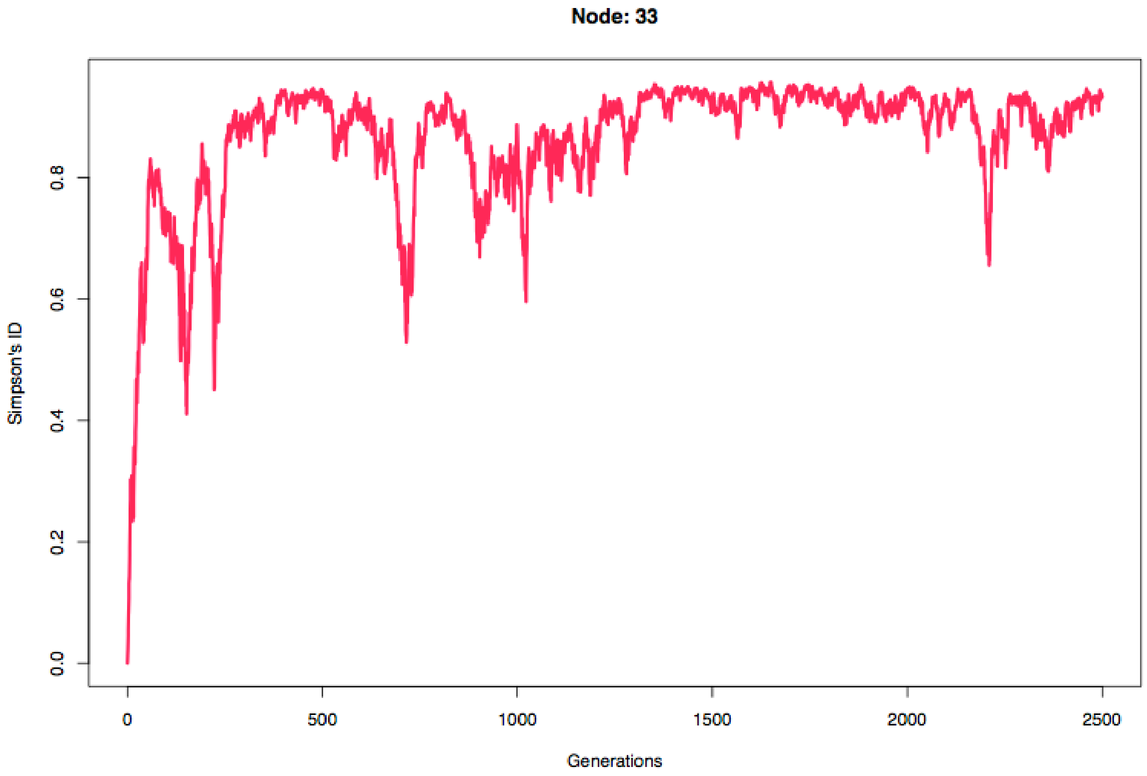

Figure 8: Karate network, diversity for node 33 with a population of 1000 individuals evolving for 2500 generations.

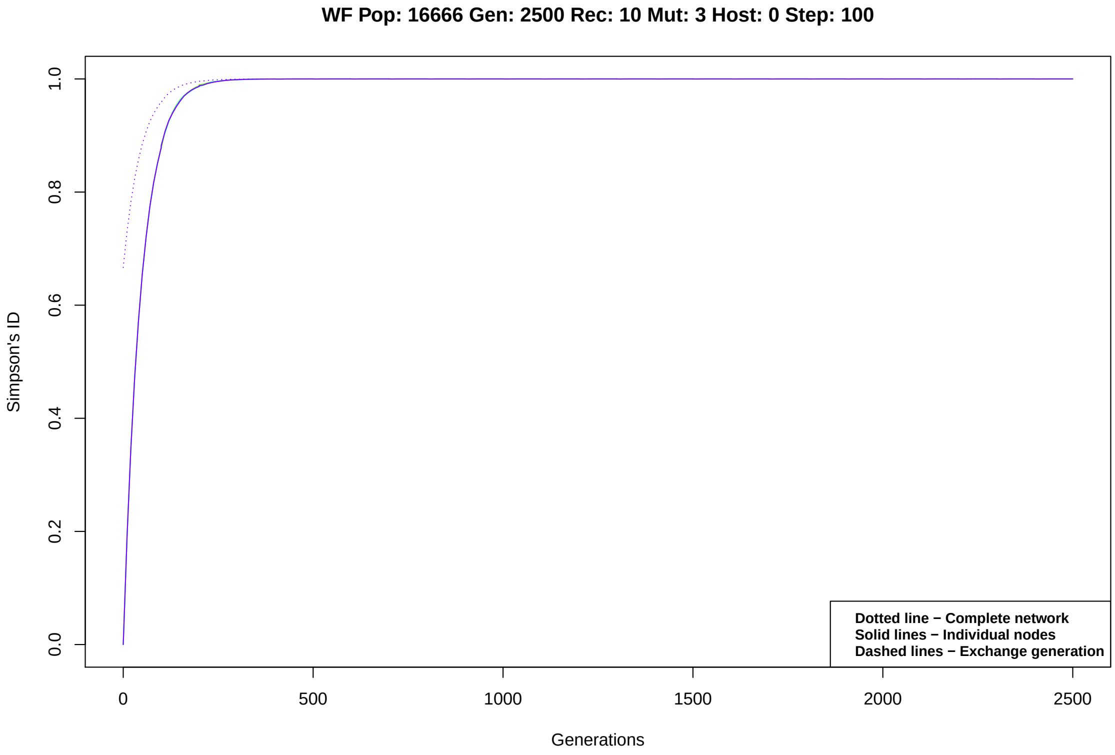

Figure 9: Stepping Stone Network SID.

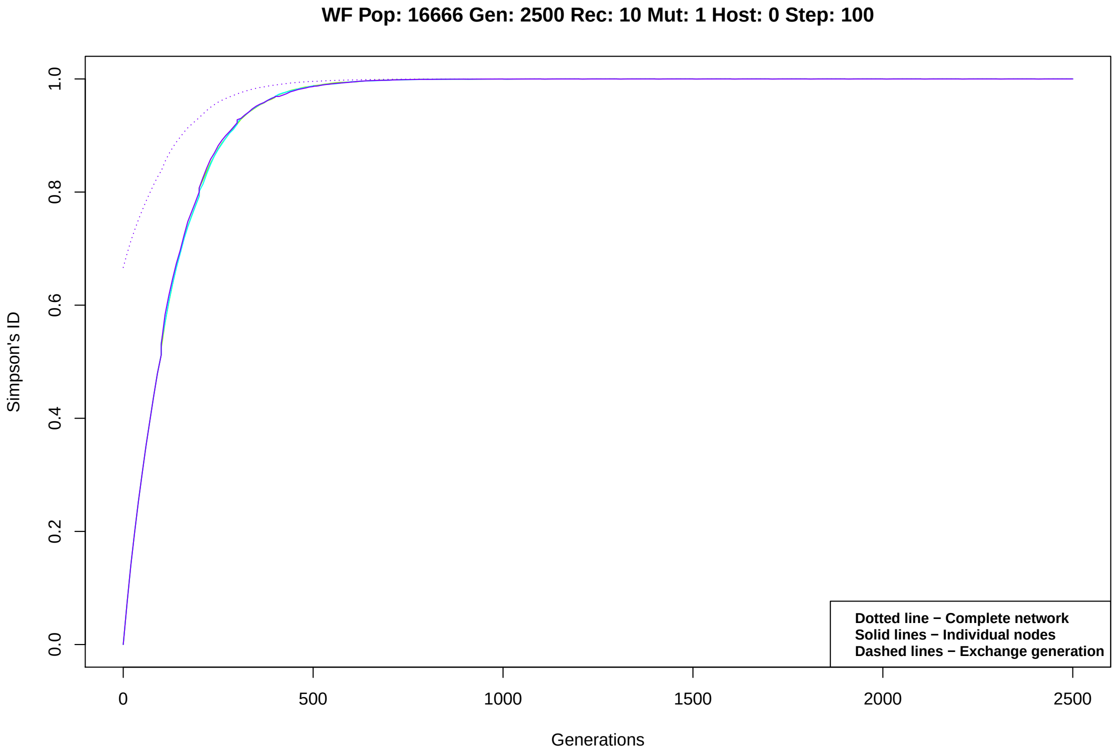

Figure 10: Stepping Stone Network SID.

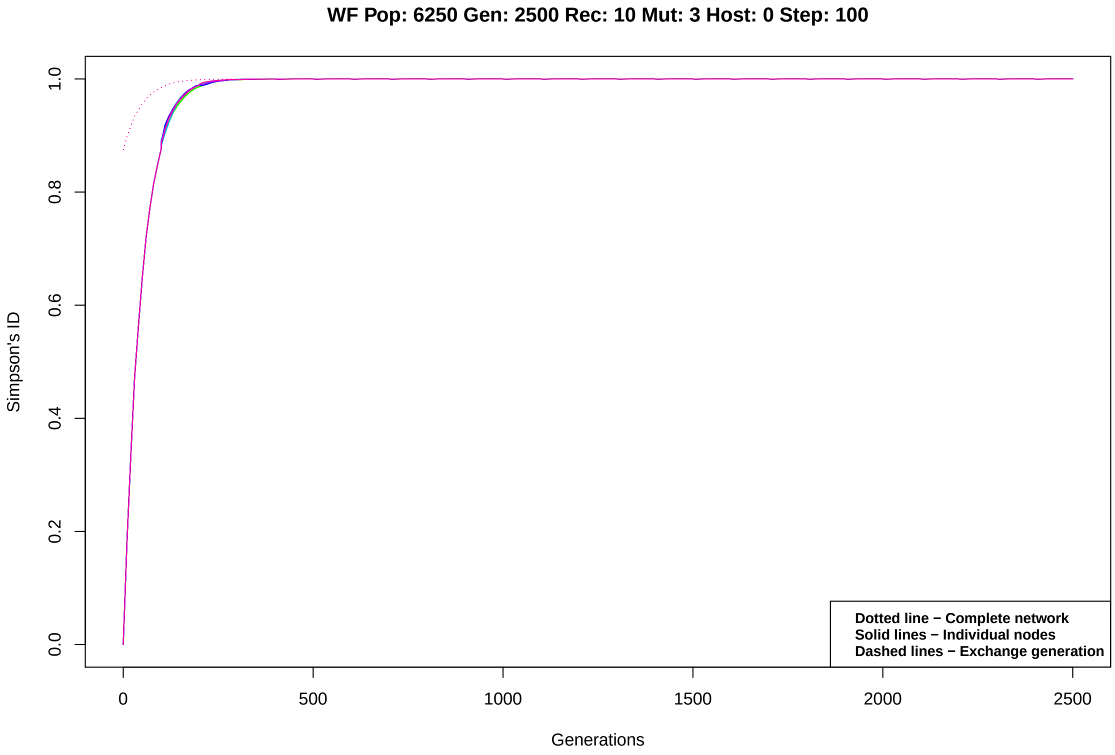

Figure 11: Cliques Network SID.

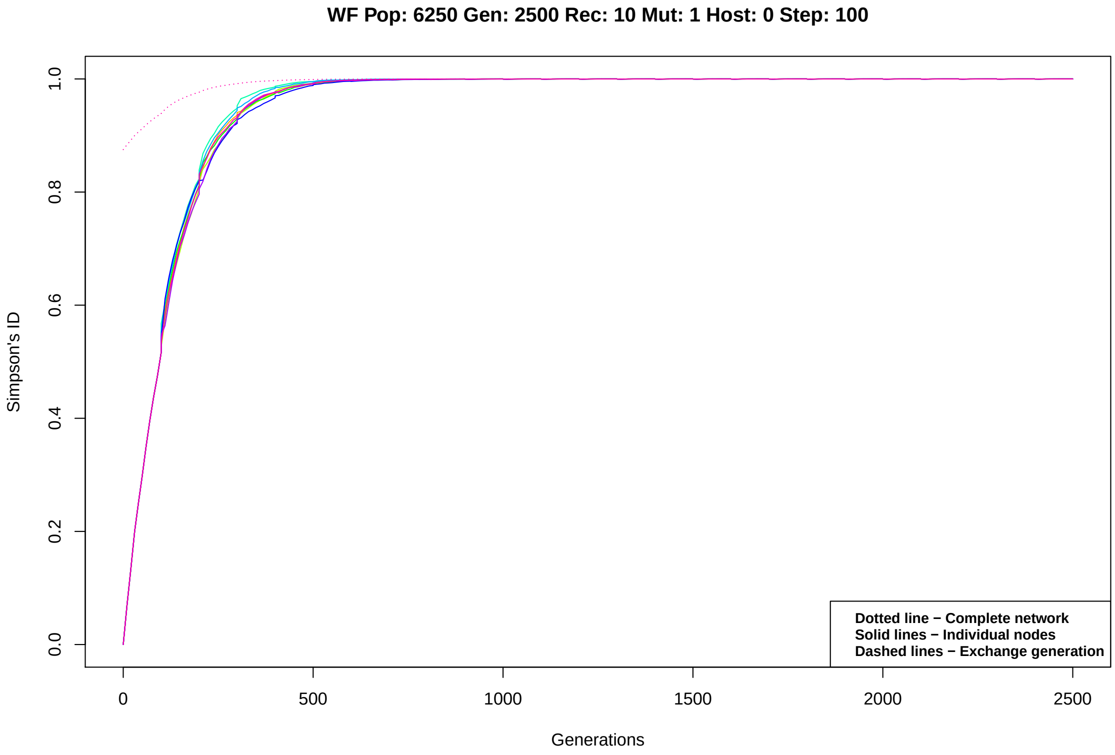

Figure 12: Cliques Network SID.

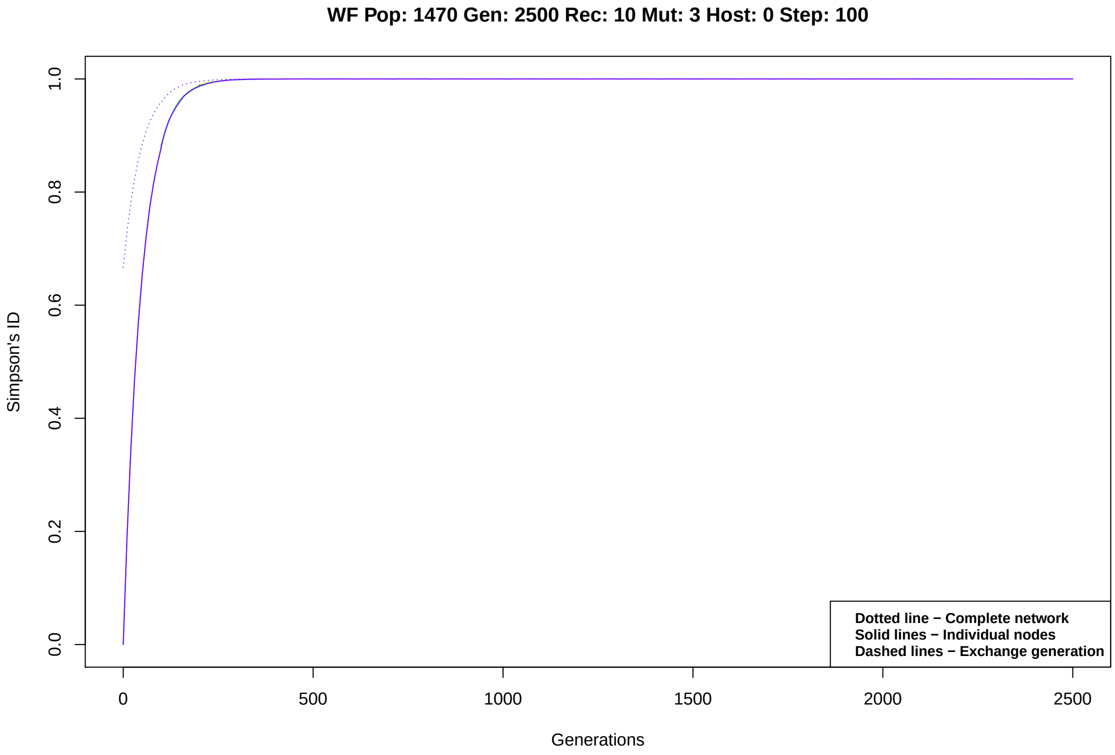

Figure 13: Karate Club Network, with bidirectional links, SID.

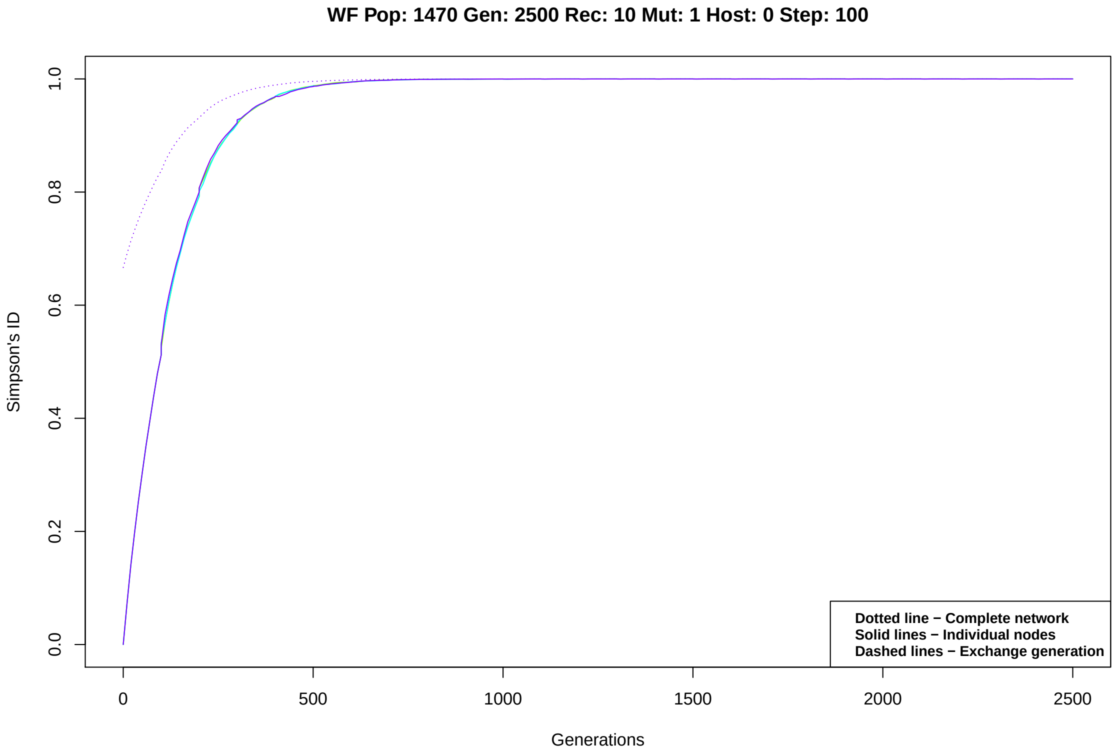

Figure 14: Karate Club Network, with bidirectional links, SID.