Simulating the Interaction of Genotype Phenotype Maps and Mutation Rates in Evolution

Sachit
20081003
IISER Pune

Thesis submitted in partial fulfilment of the requirements of Five Year BS-MS Dual Degree Program

Under the guidance of :
Dr Sutirth Dey
Department of Biology, IISER Pune.

Certificate

This is to certify that this dissertation entitled ‘Simulating the interaction of Genotype phenotype maps and mutation rates in evolution’ towards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research (IISER), Pune represents original research carried out by Sachit at IISER Pune under the supervision of Dr. Sutirth Dey, Assistant Professor, Biology Division, IISER Pune during the academic year 2012-2013.

Dr Sutirth Dey
Assistant professor
Department of biology, IISER Pune.

Declaration

I hereby declare that the matter presented in the thesis entitled ‘Simulating the interaction of Genotype phenotype maps and mutation rates in evolution' are the results of the investigations carried out by me at IISER Pune under the supervision of Dr. Sutirth Dey, assistant professor, IISER Pune and the same has not been submitted elsewhere for any other degree.

Sachit

BS-MS Dual Degree programme

IISER Pune

Abstract

In this project I have set up a framework under which several apparently disparate concepts in evolutionary biology can be analysed in a unified manner. Based on this, I implemented a software model to simulate the effects and interactions of these phenomena. Purely by changing the parameters fed into this model, it is possible to simulate phenomena like the evolution of genotype phenotype maps (GPM), epigenetics, cultural inheritance, maternal effects etc. I used this software to model the interactions of mutation rates and selection under various GP-map topographies. I find that, in line with existing theoretical results, standing genetic diversity was positively correlated with lower fitness differentials and higher mutation rates. The probability of succeeding to reach the global optimum of a rugged landscape increased with mutation rate and decreased with ruggedness. I also show that, in response to randomly changing environments, contrary to intuitive reasoning, faster fluctuations may result in *reduced* selection for mutation rates. This simulation framework, to the best of my knowledge, is the first attempt to integrate the various strands of the ongoing Extended Evolutionary Synthesis in one common theoretical framework.

Table of figures

Acknowledgements

I have to acknowledge, with fondness and gratitude, those whose touch, in my research, and in my life, made this possible.

To my mother, who has shaped
by example and through trust
what is best in me.

To Narmada Khare, Nishikant Subhedar, Shouvik Datta and D Desai, who showed me how to be, and how to love to be, a scientist.

To Sutirth, who is way more than a PI, a mentor and friend.

And to those
Who believed when I didn't
Who waited while I stumbled
Who gave what I couldn't know and didn't expect,
That I now laugh, love and learn again.

And of course all research done here is tied to Abhishek, Shraddha, Sudipta and Yashraj, who are my fellow denizens of the lab, friends, and an inseparable part of a tangled web of ideas that constitutes our research.

1. Introduction

The theory of evolution has evolved considerably since the time of Darwin and Wallace. Although Darwin propounded the twin concepts of descent with modification and natural selection, his theory lacked a credible mechanism for inheritance which prevented its widespread acceptance for the next half-a-century. The problem of inheritance was solved with the rediscovery of Mendel's work in 1900. However, instead of putting evolution in a firm footing, the rediscovery of Mendel's paper led to a bitter controversy between the so-called biometricians, who believed that continuous traits cannot have a Mendelian inheritance and the Mendelians whose stance was that the amount of variation observed in continuous traits is too less to be meaningful. This controversy raged for ~15 years before RA Fisher showed that continuous traits can have a Mendelian inheritance if one assumes them to be the additive effects of a large number of loci, each with a small individual effect (Fisher, 1919). The next three decades saw rapid progress in the areas of population genetics and quantitative genetics, which established the theoretical basis for the discipline of evolutionary biology. The findings of Darwin, Mendel, Fisher and the later workers were finally crystallized into what is known as the Modern Synthesis (Pigliucci, 2009), which is the canonical version of evolutionary biology today.

During the next 60 years, after the formulation of the Modern Synthesis (MS), tremendous progress was made in all disciplines of biology in general, and sub-organismal biology in particular (molecular biology, biochemistry, developmental biology, molecular genetics etc.). Consequently, a number of limitations of the MS became apparent. In my view, this owes a great deal to two facts that MS tries to deal with in a overly simplistic manner:

1. Heredity (defined as information passed from parent to offspring) need not just be genetic.
2. Hereditary and environmental factors interact with each other in a complicated manner.

Consequently, MS is ill-equipped to handle many questions. For example, because it has such a simplified model of how genotypes create form and function, it cannot adequately address how the modulation of phenotypes by the environment (plasticity) can affect evolution; or for that matter, how the rules of development constrain and facilitate the evolution of form or how the network properties of regulatory networks might affect their evolution. In a broader scope, the general rules governing what enables and what constrains the evolution of *new forms and functions* (evolvability and robustness) is lacking. As a consequence of its tight relationship with traditional genetics, more complicated patterns of inheritance (epigenetic or behavioural), and selection at multiple levels are hard to deal with.

The incorporation of these and other concepts into a coherent framework is called the Extended “Evolutionary Synthesis”.(Pigliucci, 2009)

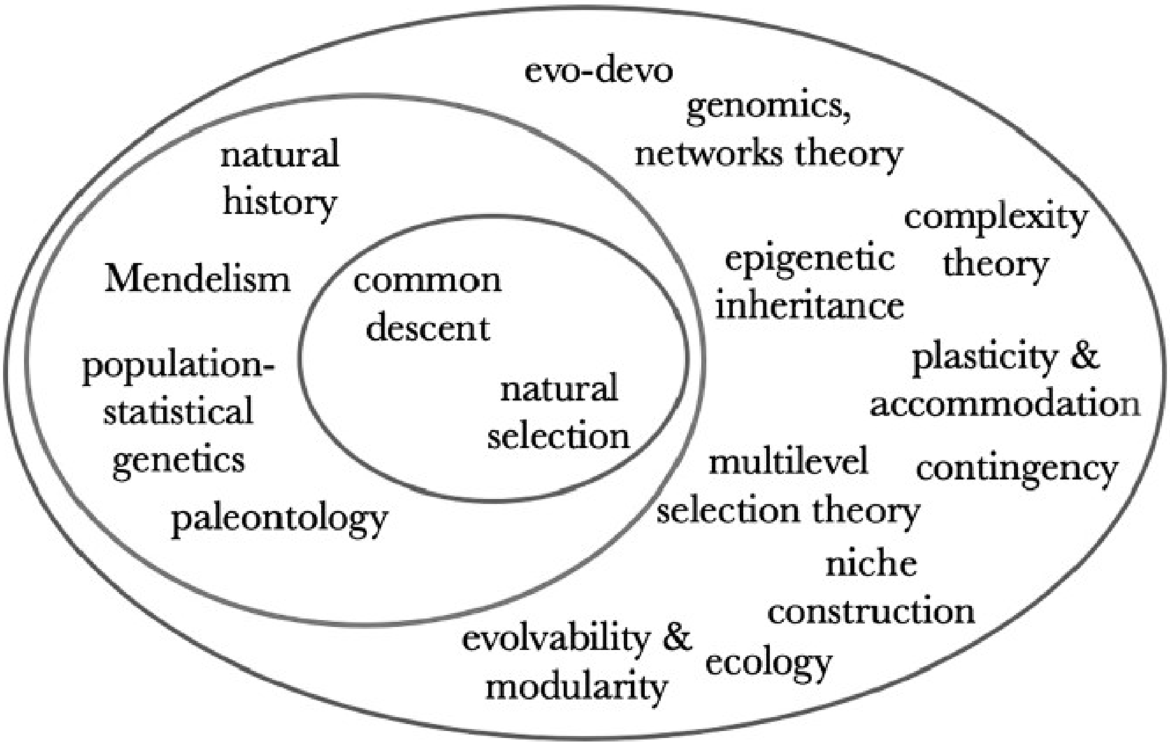

Figure 1: The Extended Evolutionary Synthesis (from (Pigliucci, 2009))

With recent advances in the understanding of how organisms develop and function (evo-devo (Mallarino and Abzhanov, 2012), metabolic networks (Basler et al., 2012), transcriptional networks (Babu et al., 2004; Crombach and Hogeweg, 2008)), and with inputs from other fields like genetic algorithms, these questions are beginning to be addressed both experimentally and theoretically. Now that computational processing power is relatively cheap, it is possible to simulate various biological and evolutionary processes in-silico. These software implementations vary in how general they are and at what scale they study biology. The scale at which the simulations are carried out determine what details must be left out or abstracted away. Simulations allow us to explore parameter ranges not observed in biology and carry out a very large number of trials in a relatively short time. This might be able to give us insights into general principles of evolution, or at least find directions along which to carry out new enquiries.

Here I have set up a theoretical framework through which several topics in the extended evolutionary synthesis can be viewed in a unified manner.

2. Materials and methods

2.1 Model

Populations change over time. These changes might be brought about by changes in the environmental conditions, the genetic make up of the individuals or even non-genetic heritable characteristics.

Many of the properties of the internal dynamics of populations can be described by the following general framework (a stochastic, individual based model with asexual reproduction):

- Organisms posses different "properties" or "traits". These traits exist in certain specific “values” or “states” in a given individual. For example, the properties: {DNA sequence at locus X, state of epigenetic switch Y, mass of stored fat} of a particular individual might be in the states { AC, methylated, 10g}.
- Some function of these values (each playing a greater or lesser role) and the organism's environment will affect the number of offspring it leaves in the next generation.
- The states of the offspring are not totally independent from their parents: the state is a (deterministic / stochastic) function of the parents states.
    - The function in question will depend on what property or trait is studied:
        - The DNA sequence AC will mutate with a fairly fixed mutation rate as function of its locus (it doesn't actually do that but let's approximate it so for a moment). What it mutates to can also be said to be fairly fixed. Given the sequence AC, we might be able to say it is most likely to mutate to AT, but it might also mutate to many other sequences each with some relative probability. Given a vast population of AC alleles, I will get the frequencies of the alleles in the next generation by multiplying the number of alleles in this generation to some "transition probability vector".
        - The epigenetic locus is a bit more complicated, what state it is in could depend strongly on its previous state but it might also

factor in the environment and genetic make-up of the parent. For example: assume that the methylated state causes the organism to express an anti-predator mechanism. Given that the organism has certain enzymes (a specific genetic background), in the presence of chemical cues from a predator (an environmental condition), methylated states in a parent might be quite likely to remain methylated. In the absence of those cues, it will be more likely that the offspring will lose its methylation. Conversely, in the first case even an unmethylated parent might preferentially give rise to a methylated offspring. These changes could be described as having a different transition probability vector for each genetic and environmental background. There are empirical examples of epigenetic transgenerational plasticity in response to predators/ herbivory (Eva Jablonka and Raz, 2009)

- The body fat of the parent might have no bearing on the body fat off the offspring, but might result in offspring that get a head start in development and begin reproducing earlier in the season. The transition probability vector of the body fat trait will vary based on a lot of other factors, but it will not vary on the parent's "state-allele for the body fat level". However, a totally different property, like start time of reproduction in the offspring's generation, might depend on the parent's state of body fat.

For a given DNA locus we will have a transition probability matrix (present allele × next allele), while for other properties ("extended loci" so to say) we will have a transition probability tensor (environment and genetic state × present state × offspring's state). In reality, even the DNA will mutate according to a tensor: certain kinds of mutations are more likely than others depending on the genetic background.

So, if the properties of an organism (some of which affect the number of offspring directly, others which affect the properties of the offspring) are considered as a "heritome" with some function relating the parent's and offspring's states, we have in hand a general lens to look at a large variety of phenomena.

To allow simulations to be carried out in a general and flexible manner, the problem was modelled using an individual based, stochastic model. There are a large number of biological processes involved in evolution and natural selection – development, plasticity, resource allocation, inter-individual interactions, fecundity, survival and mutation to name a few. These vary in detail from species to species and their relative importance in affecting evolution will again depend on the context of the system in which the species was studied. The goal of this model was to allow a large number of apparently disparate phenomena to be treated similarly in a conceptually united framework. To investigate possible general principles in evolution, these details were modeled in the following way:

- Each individual possesses a genotype or "heritome". This is a tuple of "alleles" each denoting the state or value of a particular (genetic, epigenetic, etc.) "locus" in that individual.
- The genotype of each individual is mapped to a phenotype (that may be scalar or a tuple). This mapping – the Genotype Phenotype Mapping (GPM) – may incorporate information about one or more "environmental conditions" when calculating the phenotype. The phenotype may consist of one or more traits, with each trait being set to some value or state in an individual.

GPM: P=GPM(G,E) (where G,P,E are tuples)

Genetic, developmental and physiological mechanisms that govern how an organism functions are brought together under this abstraction. For example, the concept of biological plasticity can be modelled by how the environment term is incorporated in the GPM; non-linear interactions of various loci describes the concept of epistasis when viewing the system at the level of genes, etc. Given the knowledge of how a biological process operates, the evolution of that biological process can be studied by appropriately modifying the GP map in this model. For example: given how changes to the genome interact to affect development, a mapping from the space of mutationally adjacent genotypes to phenotypes can be created; knowing the biochemical properties of the variants of the enzymes in a pathway, changes in metabolic or signalling states can be incorporated into the GPM (the genotype and phenotype vectors will connected to the concentration and activity levels of the various components of the pathway).

- The phenotype is mapped to fitness, a positive scalar.

  PFM: F=PFM(P,E) (where P,E are tuples and F is scalar)

  This mapping – the Phenotype Fitness Mapping (PFM) – defines how various traits visible to selection interact with the environment to affect the *effective* number of an individual's offspring. Factors like early life survival, competition and fecundity are all reduced to a single parameter. This approximation was made because whatever goes on between one cohort's reproduction and the reproduction of the next cohort can be boiled down to "how many individuals in this reproductive cohort come from a particular individual 'X' in the previous one?". The caveat here is that if an individual's reproductive output is a function of not just its own genotype but also of the parent's environment and genotype (transgenerational or cumulative fitness effects), this information must be added as a term in the organism's heritome.

Fitness in this particular approximation follows the rule that "the ratio of the finesses of two individuals is equal to the ratio of the expectation value of the effective number of their offspring in the next generation." (see discussion for caveats)

- When the explicit involvement of plasticity and the *mechanism* of how the phenotype interacts with environment to affect fitness are not the main topics of investigation, a convolved GFM for a given environment can be used.

  $GFM_E(G)=PFM(GPM(G))$
- A new population is created and for each new offspring, a parent is chosen such that probability of a given individual being the parent is proportional to its relative fitness.
- Along with the copying of the parental genotype, reproduction involves the mutation of the genes of the new individuals. Each locus has a "mutational map" defining the transition probabilities between its alleles. These probabilities can be scaled by an overall mutation rate which is common across alleles of an individual but may differ between individuals. These probabilities determine if an allele will mutate and what the new allele will be.

2.2 Specific Algorithmic Implementation

2.2.1 Main program Loop

- Initialization: A parameter block file is read to get the various points in parameter space over which to run the simulation. Various bookkeeping and diagnostic actions are taken like creating a copy of the exact source code of the program and all its input parameters that can be inserted into every output file so that the context of the data (which is vital for analysis) never gets separated from the data itself ensuring that the data remains reusable and verifiable indefinitely. The programme iterates over each point in parameter space and calls the main simulation function
- Simulation of a given set of parameters: Given a specific point in parameter space, these parameters are realized into the actual variables that are used by the simulation: e.g. given that the intensity of the selection is to be maintained at a certain level for a given basic GFM, the GFM is then scaled by the necessary amount to get this effect. Multiple trials are run at each point in parameter space.
- A single trial: The pseudo random number generator is reseeded at the start of every trial. Generations here are discrete (i.e. no overlap between parents and offspring). A homogeneous population of individuals with a randomly chosen genotype is used as the initial population. After this the program loops over the following steps:
    - Calculating the fitness of each individual based on their genotype (Sec 2.2.2).
    - Assigning a parent to each offspring in the next generation, based on the fitness of each parent (Sec 2.2.3).
    - Assigning a genotype to each offspring by creating mutated copies of the parental genotypes (based on the realization of the mutation probability) (Sec 2.2.4).
    - If necessary, change the environmental conditions. (implemented by the use of a different GF after a certain number of generations have passed).

2.2.2 Genotype to Fitness Map (GFM)

As the focus of the investigations here were not the evolution of the Genome-Environment-Phenotype interactions, a single convolved map for each environment that directly related points in the genotypic space to their fitness was used. By using a single combined GP-PF map, computational performance was optimized while simultaneously avoiding analytically redundant variables.

For the purpose of these simulations, the genotype space was a 2D square lattice: There were two loci with 21 alleles each. The mutational connections between the alleles resulted in the alleles forming a linear graph – every allele except the two "terminal" ones has two neighbours. A mutation will cause an allele to change to one of its neighbours.

The points along the linear graph of alleles was mapped to 21 numbers from [-1,1] that were separated by a distance of 0.1. Every point in genotypic space (g_{1i},g_{2i}) is mapped to a fitness through a function:

$F(g_{1i},g_{2i}) = f(x_{1i},x_{2i})$. Where x_{ni} is the i^{th} element of {-1,-0.9,..1}.

To generate arbitrary landscapes of varying properties (e.g. number of peaks, number of ridges etc.), $f(x_{1i},x_{2i})$ was chosen to be of the form:

$f(x_{1i},x_{2i}) = p(x_{1i},x_{2i}) * \text{SCALE} + 1$.

where,

SCALE = some constant.

$p(x,y) = [\text{Poly}_x(x)\text{Poly}_y(y)*a\ ]- b$

$\text{Poly}_x,\text{Poly}_y$ are two random 4th order polynomials

a, b are chosen such that for $x,y \in [-1,1]$ $p(x,y) \in [0,1]$;

This provides a large variety of maps differing in characteristics like the number of peaks and ridges, width of the plateaus etc. A GFM derived from a given function p corresponds to how a given set of mutationally related genotypes interact with an environment to give a set of fitnesses. For a given function "p" and its corresponding topography, the SCALE term in f(x,y) allows us to vary the magnitude of selection. i.e. different p's represent qualitatively different conditions while the same p with different SCALE factors represent

systems differing in the intensity of selection while being otherwise similar.
The calculated fitnesses were stored in a precomputed lookup table as it was found to be favourable in terms of speed vs. memory space optimization.

2.2.3 Fitness proportional reproduction

The algorithm to select parents for the next generation is a "roulette wheel selection" where the parent of a given individual is chosen stochastically with probability proportional to the relative fitness of the parents.
The naive implementation will be to make a cumulative probability distribution (which goes from 0-1 with "steps" of width proportional to the individual probabilities). Then "throw" (think of darts) a random number form 0-1 and see what step it fell on. This inverse of the cumulative distribution is best taken by a binary search. However, this takes O(OlogP) time to choose among P parents for O offspring, which does not scale well with the large population sizes I was dealing with.

The algorithm used here is a derivative of the Walker's alias table method. (Vose, 1991)

Given a set of probabilities for an arbitrary discrete distribution, we want to draw random variables with probabilities corresponding to the supplied values. If they are weights and not true probabilities such that $\Sigma_i w_i \neq 1$ they can be normalized to probabilities by dividing them by $\Sigma_i w_i$: $p_j = w_j / \Sigma_i w_i$.

To achieve this, two arrays are constructed, so that for every element e_i there is a new probability 's_i' and an index '$alias_i$'. $Alias_i$ refers to a different element of the distribution.
To generate a random variable from the distribution using the alias and s tables
1) Two uniform random numbers are drawn.

$j \in \{1,2,..N\}$

$u \in [0,1)$

2) If $u < s_j$ the element e_j is chosen else the element e_{aliasj} is chosen.

The following algorithm is applied to set up the alias and s tables:

- For all elements with $p_i < 1/N$ (where N is the number of elements) we call them 'Small' and apply the following steps:

$s_i = N*p_i$

the alias of no element refers to i

when a random variable e_x is drawn according to this algorithm,

$P(x=i) = P(j=i) * P(u<s_i)$

$= 1/N * N * p_i \because 0 \le (N * p_j) \le 1.$

$= p_i$

- For elements with $p_i > 1/N$, we initially set $s_i = N*p_i$ then recursively set some new 'Small' element k's alias ($alias_k$) to point to i and reduce s_i by $1\text{-}s_k$ till $s_i < 1$ (equivalent to $p_i < 1/N$). Then we process it as a 'Small'. So the probability of choosing an element e_i when we draw a random variable x is:

$s_i = N * p_i - \sum_k (1\text{-}s_k)$

$P(x=i)$ = P(directly choosing like a 'Small') + P(it was referred to by an alias)

$= P(j = i) * P(u < s_i) + \sum_k P($ j = k AND $u \ge s_k$)

(here the k's are all the Smalls that alias i)

$= 1/N * (N * p_i - \sum_k (1\text{-}s_k)) + \sum_k 1/N * (1\text{-}s_k)$

$= p_i$

The actual code written is in a significantly more obscure manner as it needed to be optimized for space and time. The main differences in the actual code are:

- The condition for $\Sigma_i p_i =1$ is relaxed.

Finesses are directly input and used without normalization.

- The variables s_i are redefined as: $s_i = p_i$
- A new variable threshold is defined: threshold = $\Sigma_i p_i/N$
- Condition for a p_i to be 'Small' is redefined as: p<threshold :

$p_{normalized} < 1/N \Leftrightarrow p < \Sigma_i p_i /N$

- When recalculating s for large p: s_i -= threshold - s_k
- Condition for choosing between j and alias_j:

$$u*\text{threshold}<s_i: u < N * p_{\text{normalized}} \Leftrightarrow u * \Sigma_i p_i/N < p_{\text{non normalized}}$$

2.2.4 Mutation

For an allele at a given locus, we create a list of all its mutationally adjacent neighbours and a list of weights proportional to the transition probabilities from the given allele to its neighbours. These two components can be described as follows:

- Neighbours = $\{a_1,a_2,a_3,...\}$ where a_i is the identifier for the i^{th} neighbour that is one mutation away from the given allele.
- TransWeights=$\{Nw_1,Nw_2,Nw_3,...\}$ where w_i is the transition weight to the i^{th} neighbour and N is the number of neighbours.

To mutate a given gene (an allele at a locus), we draw two uniform random numbers:

$j \in \{1,2,..N\}$

$u \in [0,1)$

If u<TransWeight_j·MutationRate the locus mutates to its j^{th} neighbour, else it does not mutate. This is repeated using the respective tables for every locus of every individual.

Proof of correctness / explanation of working:

The probability that a gene will mutate to its j^{th} neighbour is

$$p_j=p(\text{choosing slot } j = j) * p(u < (N * w_j *\text{MutationRate}))$$

$$= 1/N * N * w_j *\text{MutationRate}$$

$$= w_j *\text{MutationRate}$$

(The only constraint on w_j is that $0 \le (N * w_j *\text{MutationRate}) \le 1$.)

the probability of a mutation occurring is

$$p_{\text{mutates}}=\Sigma_j p_j$$

$$=\text{MutationRate} * \Sigma_j w_j$$

For the purpose of this simulation, the genotype had two loci that affected fitness. Both of these had a mutational map of 21 alleles connected linearly with a terminal node at each end, i.e. every allele except the 1st and the 21st had two neighbours. The transition probability to both neighbours was equal.

When the mutation rate was allowed to evolve, a similar linear map of 6 mutational alleles was used for the locus that determined the organism's genome wide mutation rate. A look-up table was used to translate the allele present to a floating point rate.

2.3 Investigation

The implemented framework was tuned to study mutation rates in evolution placed in the context of different GF maps that abstracted various environmental conditions and genetic constraints.

The three main areas of interest were:

- Simulating the presence of standing neutral variation in the presence of stabilizing selection (mutation-selection balance)
- Simulating the relationship of mutation rates and selection intensity in governing the success of evolution in reaching optima / the ability of populations to execute valley crossings in rugged fitness landscapes.
- Simulating the effects of fluctuating environments on the evolution of mutation rates.

The first two areas were simulated using the same model:

Since the mutation rate was an independent driving variable in the system, it was held fixed across a population, across generations, in a given simulation. The simulation was conducted for the following parameters in a factorial design:

- 10 maps of varying properties.
- Three different selective gradients: with the minimal fitness being 1 in all cases and the maximum fitness being 1.1, 2 or 11.
- Three different mutation rates
- Every point in parameter space was repeatedly simulated with 14 different starting populations with 96 trials for each starting condition.

The effects of fluctuating environments on the evolution of mutation rates was studied by using a slightly modified model: mutation rates were encoded in the individual organisms and allowed to freely evolve. The mutation rate of an organism could vary between 10^{-1},10^{-2},...10^{-6}. A change in environment was effected every X generations by changing the GF map in use. The simulations were carried out in this factorial design:

- 3 classes of GF Maps corresponding to different systems with different evolutionary constraints: maps with 1, 3 and 9 maxima.
- 5 treatments each corresponding to a different sequence of maps of a given class.
- 5 Population sizes varying from 50 to 10000
- The time period between changes in environment set to 10, 20, 30, 50, 100, 150, 300 or 1000 generations.
- 3 Selective gradients corresponding to maximum finesses 1.1, 2 or 11 (minimum finesses being 1).
- Every point in parameter space was simulated with 96 different starting populations.

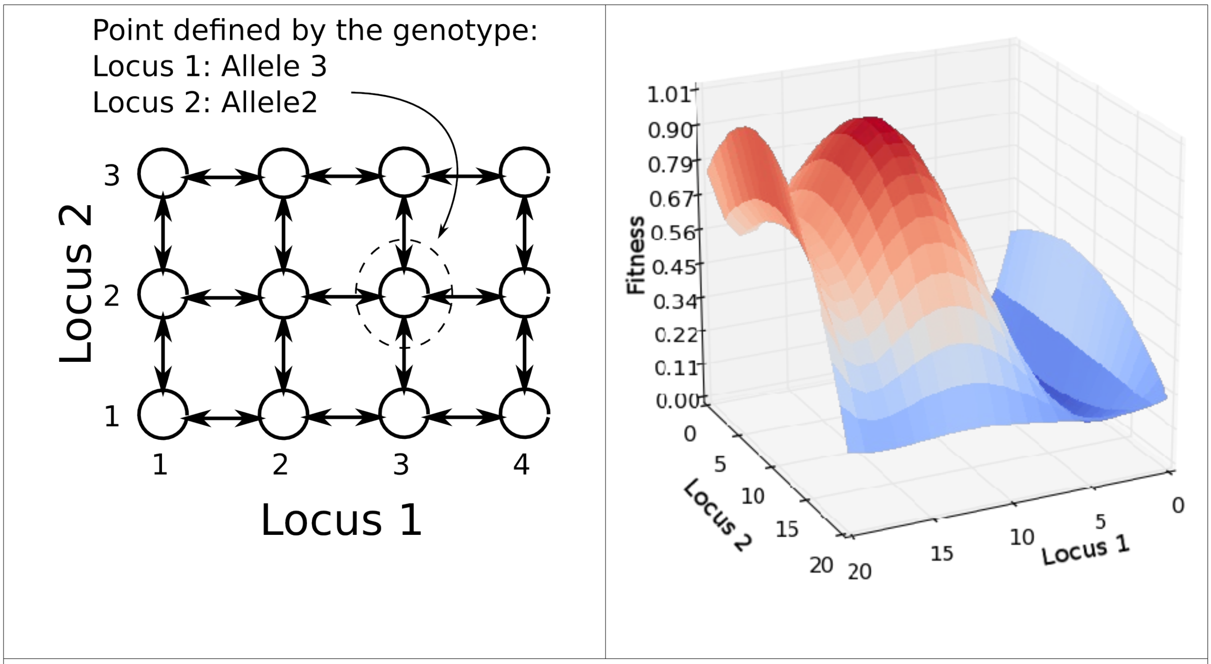

Figure 2: (A) A small subset of the mutation map (i.e. the genotype space) used in this simulation. (B) An example GF map.

3. Results and Discussion

3.1 Simulation results

3.1.1 Mutation selection balance – standing genetic variation in a population in the presence of selection

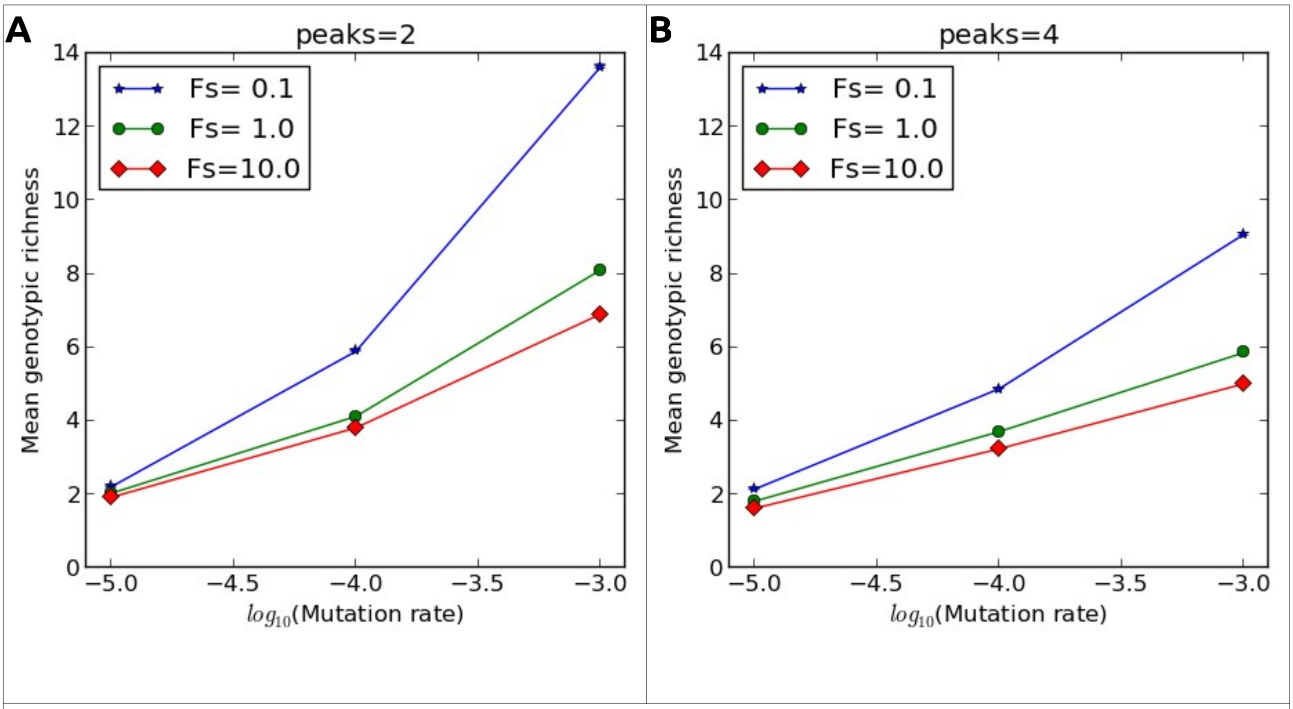

Figure 3: Mutation - selection balance and genotypic richness. Mean genotypic richness (i.e. the number of unique genotypes in a population) after evolution for 10000 generations plotted vs. the mutation rate (10^{-5} to 10^{-3} mutations per locus per individual per generation). Simulations were repeated with different levels of fitness scaling for each GF map. (A) Data from GF maps with 2 peaks (B) Data from GF maps with 4 peaks. Fs = Fitness Scale (in a given map, Fitness_Scale = Max_fitness/Min_fitness -1). It can be seen that standing diversity increases with mutation rate and decreases with selection gradient.

The number of unique genotypes present at the end of each trial was counted and plotted against the fitness scaling factor and the mutation rates. The following results were observed (Figure 3):

- With decreasing mutation rates, the standing genetic variation in a population went down (due to the decreased production of new genotypes per generation).

- Standing genetic variation was reduced by increased fitness scale coefficients (due to increased efficiency of removal of less-fit genotypes).
- For very low mutation rates, the selection gradient is no longer a determining factor in the level of variation – so few new variants are produced that drift and very weak selection are sufficient to remove them.
- Maps with more densely spaced optima (narrower “peaks” , Figure. 3B) had less standing variation at the peak that was populated compared to those with fewer peaks (Figure. 3A).

This model demonstrates the basic features of the mutation selection balance, which is one of the properties any evolutionary model incorporating natural selection should show. These results indicated that the our model leads to qualitatively similar predictions as standard population genetics theory (Falconer and Mackay, 1996), which reassured us in terms of using valid assumptions for the framework and served as a very indirect check on implementation bugs. The exact shape of the curves will depend on the dimensionality of the genotypic system (higher dimensional surfaces amplify the effects of mutation rate) and the population size (smaller population sizes have lower allelic richness).

3.1.2 Traversal of rugged fitness landscapes

Populations were allowed to evolve under different conditions of fitness scaling and mutation rate in GF maps that had 2 or 4 fitness optima. Populations that get stuck in local optima sometimes escape them and reach higher fitness optima. The frequency at which this happens can be very important in biological systems, as typical biological GF-maps are very complicated.

As a measure of how successfully populations cross “valleys” of low-fitness intermediate genotypes, the fraction of the populations that managed to reach the global optimum was measured.

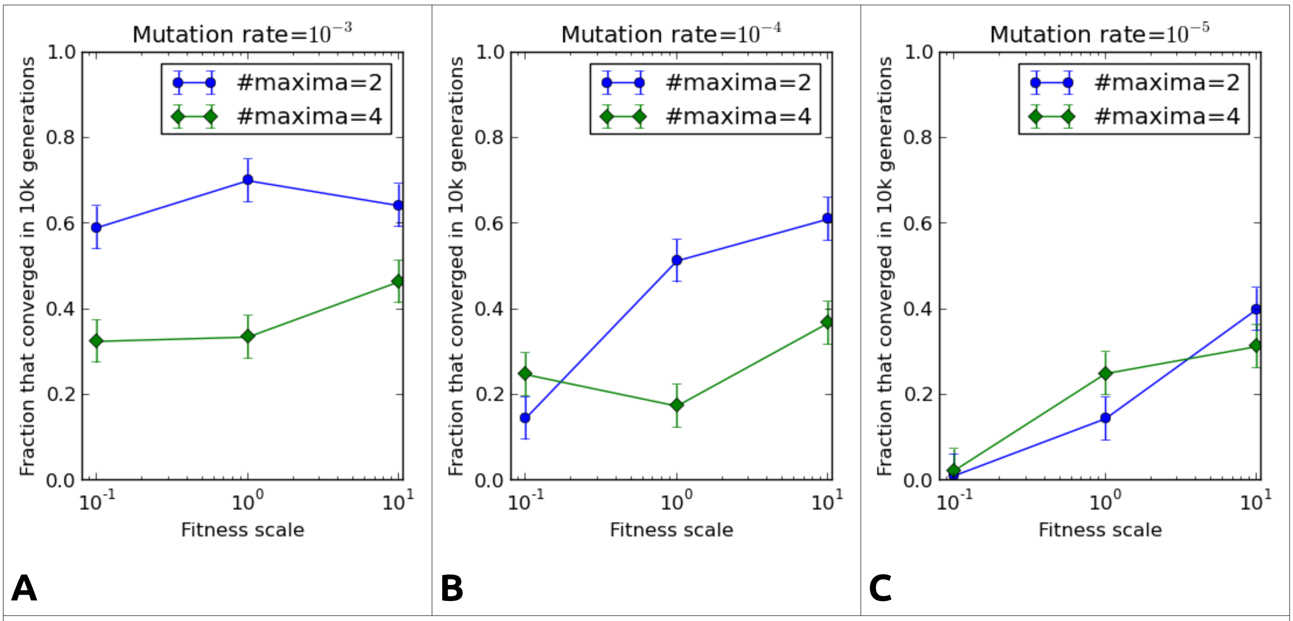

Figure 4: Traversal of rugged fitness landscapes. Fraction of trials that converged to the global optimum after 10000 generations vs. the fitness scaling. (A-C) results for the simulation at mutation rates 10^{-3} to 10^{-5}. As sample size* fraction >10 in all cases, the normal approximation of binomial standard error was used. i.e. a list of 1's and 0's corresponding to the converged and non-converged states was generated and it was treated as normal distribution.

The following trends were observed:

- Maps with more peaks are more likely to lead to populations getting stuck (Figure 4, all panels). This result is intuitively expected. However, for very low mutation rates, mutation rates are the limiting factor in evolution, not the topography.
- As might be expected, higher mutation rates allowed larger fraction of the trials to converge to the global optima.
- At high enough mutation rates, the influence of fitness scaling wanes (Figure. 4A), and that of GFM topographies increases (Figure 4A vs 4C).
- Greater fitness-scale factors lead to higher probabilities of reaching the global optimum. This could appear counter-intuitive, because greater penalties for entering valleys should result in more, not less, populations being trapped. The reason for this result is that for the lowest fitness scales, local gradients are insufficient to cause any significant evolution. The probability of reaching the global maximum also seems to depend on the presence of ridges in the landscape that serve as attractors or almost-neutral bridges to global maxima. Individuals entering these

attractors are more likely to spread if the gradients are steeper (see Figure 5, 6). As these figures demonstrate, the probability of convergence is not smooth with distance when a few random start points are chosen, suggesting the presence of location-specific attractors. At very low levels of positive selection, these ridges will be effectively neutral, and there is an “entropy barrier” which holds the population at evolutionary stasis (Vannimwegen and Crutchfield, 2000). The behaviour of random walks in real multidimensional GF maps can be quite different from that of the simple 2D lattice used here (Østman and Adami, 2013; Pigliucci, 2010). This is primarily mediated by the presence of high dimensional ridges that bypass valleys. The importance of the overall selection gradient in crossing almost-neutral zones needs to be looked into further for higher dimensions.

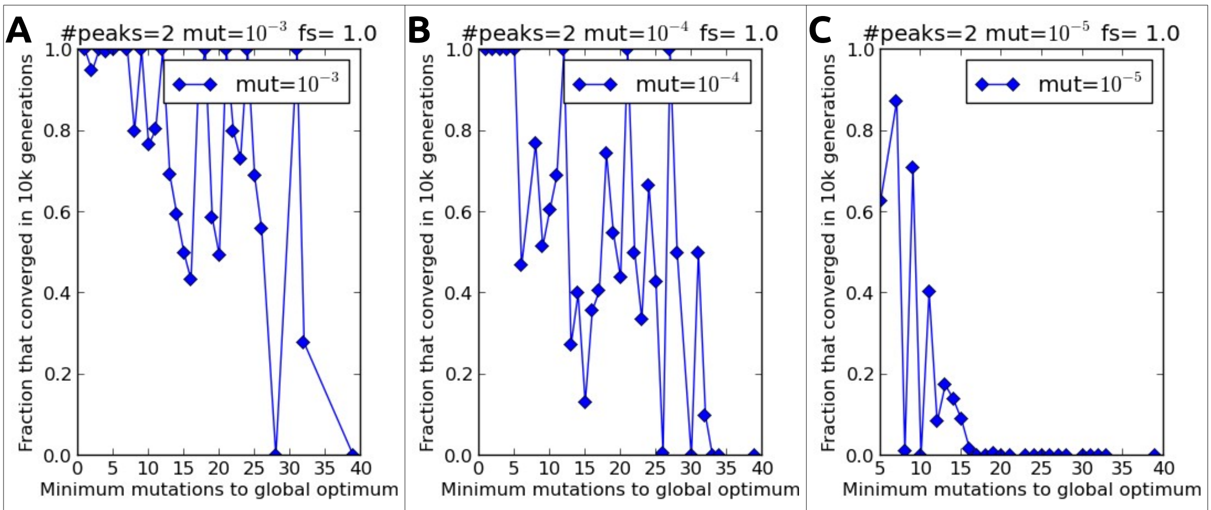

Figure 5: Interaction of mutation rates with the traversal of rugged GF landscapes. Fraction of trials that traverse the landscape and reach the global optimum as a function of the minimum number of the mutational steps between the starting point and the global maximum. (A) mutation rate 10^{-3} (B) mutation rate 10^{-4} (C) mutation rate 10^{-5} . Distance of the start point from the global optimum does not have a smooth curve relating it to the probability of convergence. Certain locations, though further away from the global optimum, result in greater convergence, probably because they are closer to ridges that connect them to the global maximum.

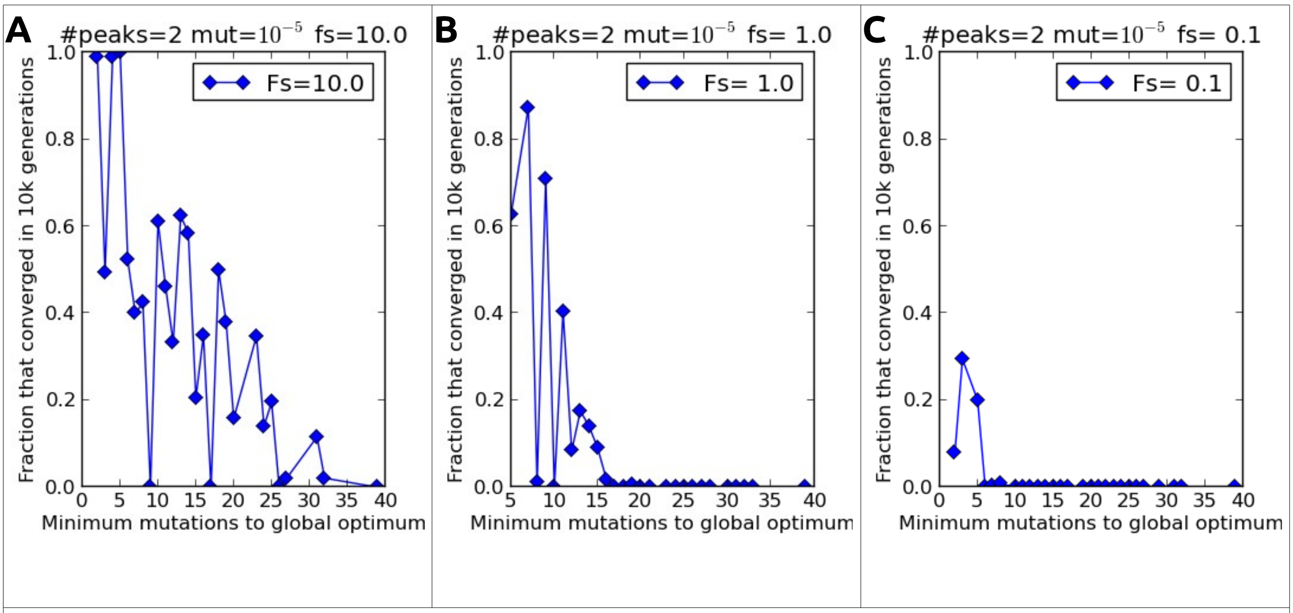

Figure 6: Interaction of selection scaling with the traversal of rugged GF landscapes. (A) Fitness scale=10 (max. differential fitness ratio=11) (B) Fitness scale=1 (max. differential fitness ratio=2) (C) Fitness scale=0.1 (max. differential fitness ratio=1.1). See Figure 5 for further explanations.

3.1.3 Effect of fluctuation selection on the evolution of mutation rates

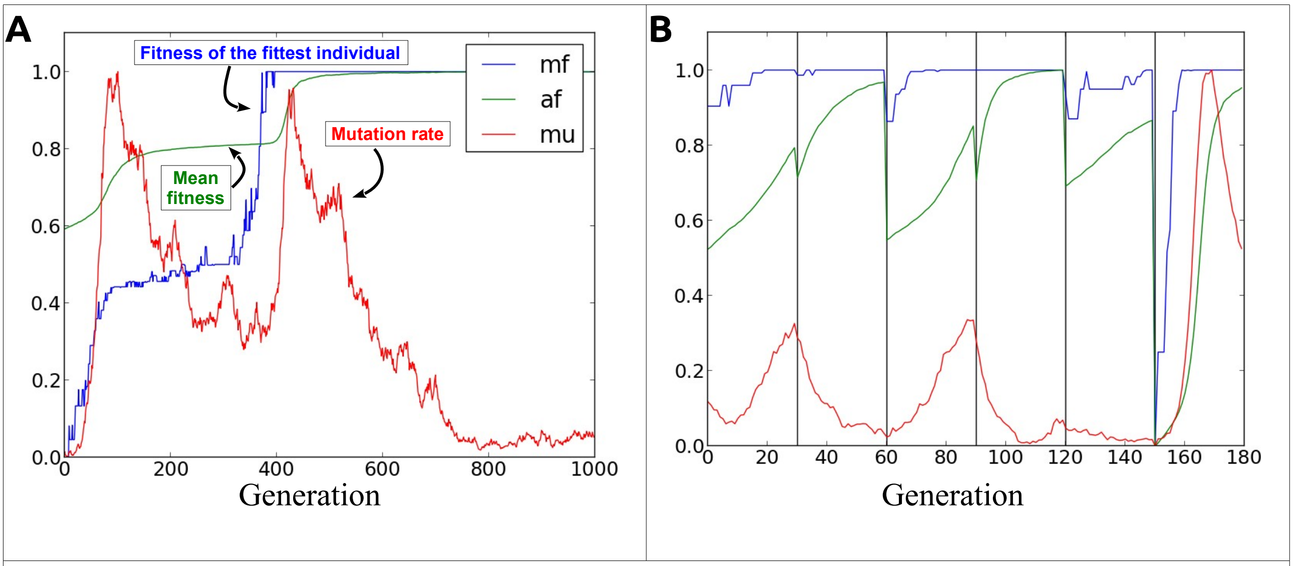

Figure 7: Effects of fluctuating selection on evolution of mutation rates and fitness. All values were scaled to 0-1 by normalization. (A) An atypical set extracted from a simulation of fluctuating selection with the time period of fluctuation=1000 generations. This data series is atypical in that evolution stalls for a long time at a point before the population resumes evolving. (B) 180 generations of data (legends omitted for clarity. Same scheme as A) for a fluctuation time period = 30 generations. Each change in environment is marked with a vertical black line.

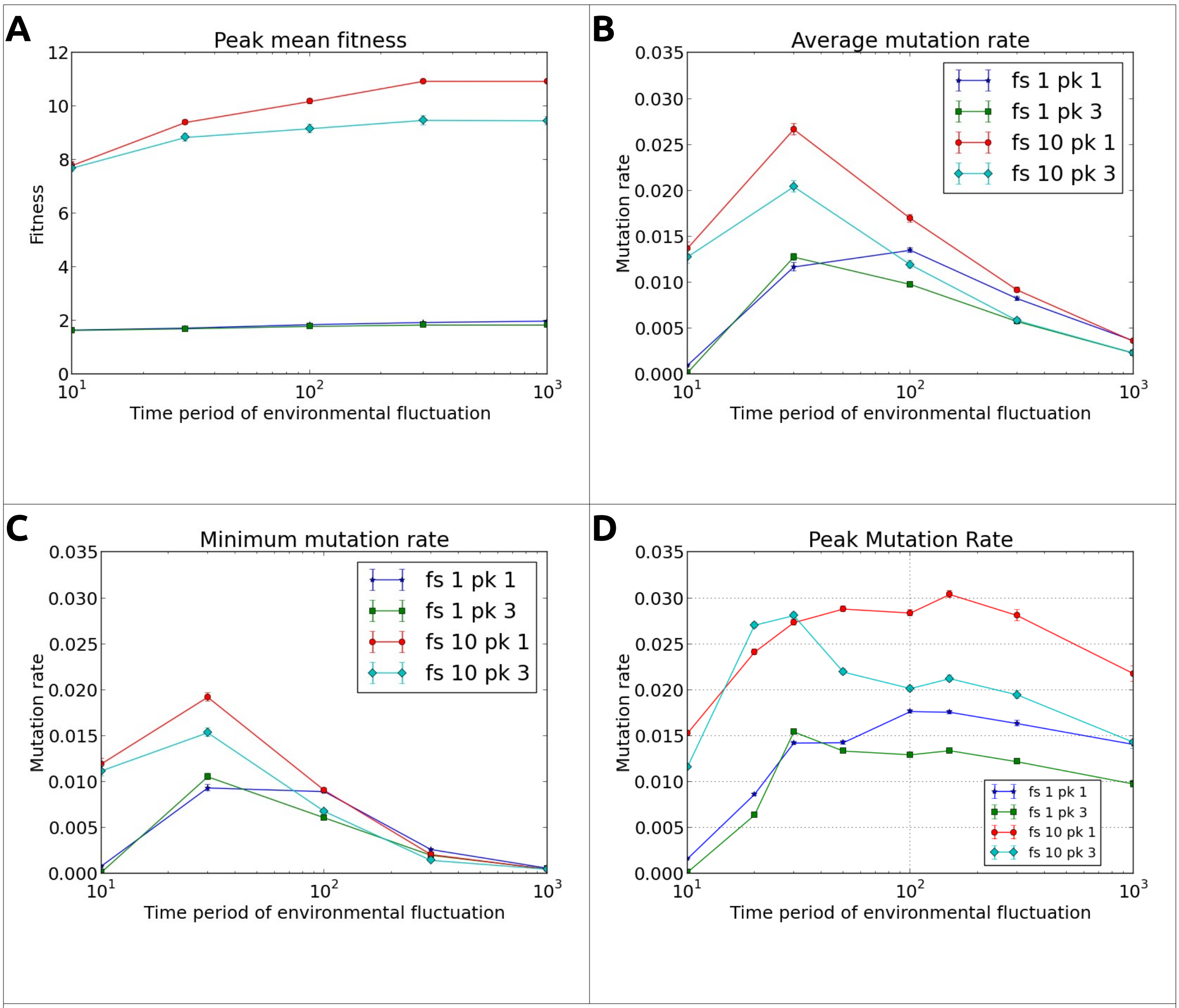

Figure 8: Evolution of mutation rate with fluctuation frequency under different conditions of selection strength (fs) and map topography (pk). (A) The maximum fitness a population achieved within the period it was allowed to evolve before the environment was changed. (B) Average mutation rate of the population (C) Minimum of the population average mutation rate achieved within the time it is is given to evolve to an environment. (D) Interaction of selection gradient (fs) and the number of peaks (pk) with fluctuation frequency. Note: Legends are dropped in some panels for clarity; the same labelling scheme is used for all plots.

The main results observed here are:

1. Due to the presence of a mutational bias against high mutation rates, irrespective of both topography and selection strengths, the mutation rate decays to the same value over time (Figure 8 C, time period 300 and 1000).
2. Even in the presence of this bias against higher mutation rates, it is observed that the mutation rate rises while the population is evolving (Figure 7A, 7B). This proves that there is a significant 2nd order positive selection on mutation rates, brought about by the hitchhiking of the mutator allele along with newly created beneficial variants. This is a case of selection for evolvability, which has been empirically demonstrated in *E. coli* populations (Shaver et al., 2002).
3. In the case of rapid environmental fluctuation, slower fluctuations (time period = 30) have a consistently and significantly larger mutation rate compared the more rapid fluctuations (Figure 8D).
4. For a large number of conditions, (T=30 - 1000) the maximum achieved mutation rates are very close for a given map topography and selection scale (Figure 8D).
5. Higher selection scales lead to higher mutation-rates.
6. Smoother maps have higher rates of mutation (9 peak maps have even lower rates than 3 peak maps; data not shown).

3.1.4 Discussion and analysis of results

It was observed that when the gaps between environment changes was large, in the order of 300 to 1000 generations, the final mutation rate did not change with the selective pressure or scale. This suggested that evolution was being driven primarily by forces other than natural selection. The nature of the curves resembled exponentials (Figure 7). Exponentials are generated in nature when the rate of decay (dx/dt) is proportional to the present value. Examining the model closely for a force that would directly (without intermediate steps) reduce mutation rates revealed the source of the anomaly as a mutation bias.

This bias happens because the transition probability vector for the mutational alleles had *equal* weights for mutations to either neighbour. However, this unbiased vector generates a mutational bias of its own (only) when it occurs on a locus that affects mutation. Consider four alleles with mutation rates 4, 3, 2 and 1 that lie somewhere in a chain of such alleles. Assume that the biases to mutate to the left neighbour are the same as those to mutate to the right neighbour, i.e., the allele with mutation rate 2 is equally likely to mutate to 3 or 1. Similarly, 3 is equally likely to mutate to 4 or 2. However, because 3 mutates at a higher rate than 2, 3 is more likely to mutate into 2 than the reverse happening. This results in the mutational locus steadily moving down the linear chain of alleles till the population is mostly composed of the lowest mutating allele. This manner of evolution (mutational bias) is fairly important for certain genomic regions (Ellegren, 2000; Marais, 2003), but unfortunately, that is not what I was trying to study. To correct against this, the mutation transitions must be counter-biased to give an unbiased mutational landscape. As a consequence of this bias, I was unable to infer anything about the existence of a second order selection against high mutation rates in this system.

Interestingly however, mutation bias provides two properties that are potentially very useful:

- In the absence of any selection, through mutation bias, the mutation rate is rapidly brought back to a lower value **without** affecting the allelic frequencies of other linked loci (selection at any locus will distort the allelic composition of other linked loci).
- The “force” driving the change in gene frequencies should increase in a very predictable manner as the population mutation-rate increases. The selection “force” on mutation can potentially be measured against this force. The peak in the mean mutation rate occurs where the forces increasing and reducing mutation rate balance each other out. The two forces in this case are the selection for evolvability that is acting to increase the mutation rate and the mutational bias that is acting to reduce it. As is shown below, the value of the peak mutation rate can be used as a proxy for the selection differential on the mutation-rate.

d Mutation-Rate/dt =0 → Mutational bias =selective pressure

Mutation bias ~ Mutation-rate

(the data showing the smooth relationship between the two is not shown).

selective pressure ~ Mutation-rate$_{\text{maximum}}$

Due to these advantages, the bias was preserved in later simulations.

Mutators spread when they are linked to high fitness alleles that sweep through the population. Higher mutation rates do not spread if the next beneficial mutation arises before a sweep completes, especially if it arises in a lineage that is not currently taking over the population. This “clonal interference” plays a very important role in limiting the advantage of ever increasing mutation-rates (G et al., 1999). If the new beneficial mutation arises in the population before the previous beneficial allele is fixed, it will end up competing with it instead of adding to the fitness. The optimal mutation rate will be the one where a new allele is created just as the previous one approaches fixation, ensuring that the effects of the two add together for an even fitter phenotype. Faster mutation-rates can't capitalize on this (their new beneficial alleles appear before the previous allele is fixed and are “wasted” most of the time because they end up competing with the existing clones)

Clonal interference might play a crucial part in the limitation of mutation-rates in this model too.

1. Higher selective gradients lead to the faster sweeps of beneficial alleles. Mutations can now occur at a higher frequency. This explains the observed trends in mutation-rates. Also, the mutation rate for many fluctuation frequencies is similar for a given set of conditions, because thanks to clonal interference, mutations are no longer a limiting factor for the maximum rate of evolution.
2. The presence of multiple peaks in the landscapes may result in the rate of evolution going down during periods when it is stuck on a lower peak or on a low slope ridge. This other factor limiting the maximum rate of evolution will result in clonal interference reducing the benefit of higher mutation rates.

As mutator alleles gain their fitness by hitch-hiking along with the beneficial mutations, the difference in time for which their mutant clones can expand will affect their fitness. When mutation rates are the sole limiting factor, the difference in fitness between two mutator alleles will increase with the difference in the mean time it takes each of them to come up with the next beneficial mutation. Assuming exponential growth in the early expansion of two identical clones with a relative-fitness-over-the-ancestor = r, the relative fitness between the two identical clones that arise 'g' generations apart will be r^g . If only a limited window consisting of the early part of an evolutionary time series is taken, the mean time by which the lower mutator lags behind the faster mutator will also be reduced (we are constraining how late a mutation can arise, and thereby also limiting the maximum difference in initial mutation time). Therefore, when only a small early window is considered, the advantage of higher mutation rates is lost. This could be behind the observed increase in mutation rate with lowered frequency of fluctuation.

3.2 Discussion of model

The mutation bias results in the population gradually being dominated by the lowest mutation rate. However, this is completely opposite to what we know about how mutator alleles work. Most mutator alleles arise due to loss-of-function mutations which can happen at many positions in the mutator sequence, which is why it happens at a relatively high rate (Denamur and Matic, 2006). However, gaining back that function requires a back mutation exactly at the point where the previous mutation has happened or many compensatory mutations (Burch and Chao, 1999), which clearly happens with a much lower probability and may not even restore full functionality (Wielgoss et al., 2013). To correct against this bias in the model, the mutation transitions must be asymmetric with higher probability of mutating into alleles with high mutation rates. As a consequence of this bias I was unable to infer anything about the existence of a second order selection against high mutation rates in this system, which is what I am planning to do next.

3.2.1 General comments about the simulation framework

The chief advantage of the simulation framework is that it can be used to investigate many different evolutionary phenomena and the interactions between them under one common rubric. i.e. data from many fields (e.g. evo devo, transcriptional networks) can be handled as part of the GF maps while phenomena like epigenetics can be inserted into the mutation transition weights.

3.2.2 The fitness landscape: the GFM and the mutational network

As the evolution of environmental interactions was not under investigation, a single convolved map was used, i.e. the GP and PF maps were collapsed to a single G-F map. The characteristics of the map were chosen to be random topographies in an attempt to extract general principles.

The chief advantage of this approach lies in that the evolution of the GFM itself can be studied by allowing its parameters themselves to be loci on the heritome.

3.2.3 Strengths and weaknesses of fitness scheme

Pros:

- It is a highly general description of what fitness is. Suitable GFMs should easily describe many biological cases, allowing the clean use of this implementation of fitness and differential survival.
- Variations in inter-generational population size can happen at this stage without any modifications by just choosing the number of new offspring allowed at each step (fitness is just relative representation here). This allows the model to easily capture the effect of population dynamics on evolution and the evolution of population dynamics (if the population size is a function of the P vector).
- Life history: Simulation of overlapping generations (as opposed to the discrete generations used here) can be easily achieved by combining the new generation chosen by this scheme with a subset of the parental generation. The only change required is to define a scheme of choosing the surviving parents: random death, age, etc. This will require adding a new property to the set of properties of each individuals.

Caveats:

- It will be difficult to capture certain life history strategies without modifying the scheme of selection. For example, lets consider the case where the possible representation in the next generation is a saturating function. Assume that in the offspring generation, the mean of the absolute representation of an individual with relative fitness f is:

  $N_{offspring}$ =max(Population_Size$_{new}$ * f , o)

  where o is the maximum number of offspring the organism can produce and f is the relative fitness.

  This is best understood with a concrete example: consider an organism that produces a maximum of 3 offspring and of its offspring have to compete for a limiting resource essential to survival. If a single individual of a very competitive type is placed in a small population of weaker individuals against whom it always wins, then irrespective of the relative fitness, that individual can only have 3 offspring that represent it in the next reproductive cohort. Now consider another phenotype that is 20% more likely to win fights against the first type (it also produces 3 offspring). A single individual of this maximum fitness phenotype cannot be distinguished from the middle fitness phenotype in a population of the minimum fitness phenotype: they both will produce the maximum of 3 offspring. However, when the two fitter types are mixed, the difference becomes visible. The presently used scheme (N1/N2 =F1/F2 where N is the number offspring and F is fitness) cannot describe this system.

- The present study assumes that the fitness of an individual does not depend on the properties of its parents (except the static genome), which is an accurate description for many traits. When this assumption is not true (e.g. maternal effect (Galloway, 2005), trans-generational plasticity (Eva Jablonka and Raz, 2009) etc.) the extended heritome discussed previously could be easily implemented. This would only require increasing the dimensions of the mutational map by one (also see the next section).

3.2.4 Advantages and biological relevance of mutational implementation

This system allows enormous flexibility in the modelling of biological phenomena.

- In the present version of the model, we used a square lattice mutational network. However, true mutational networks are likely to be high dimensional and the program has been implemented in a way that it can be extended to mutational-neighbourhood-network of arbitrary dimensions and topology. The same code is reused for all networks. However, the relatively smooth, low dimensional map used here may be thought of as an approximation of the coarse grained view of genetic components, where the interaction of many small factors result in a overall smooth behaviour.

- The present implementation allows fine-grained control of the mutation rate at many levels – population-wide, genome wide, locus specific and even allele specific. Modifications to each level can be made in a simple and consistent manner by the user.
    - Different genes (gene implying allele at a locus) can have different relative mutation rates controlled by the respective transition-weight vectors for each gene. The magnitude of $\Sigma_j w_j$ can vary between genes, and the total mutation probability of a gene is MutationRate * $\Sigma_j w_j$. This allows the biological phenomena of mutational hotspots on the genome, and alleles that have higher mutation rates, to be captured. Certain regions of the genome have higher than normal mutation rates and biases (Green et al., 2003), while in some cases the rate of mutation depends on the allele in question (Schlötterer et al., 1998; Yu et al., 1991).
    - The "MutationRate" scaling parameter can be varied on a per-organism and a population wide manner. This allows us to capture biological phenomena like the presence of environmental mutagens (through the population-wide scaling of the mutation rate), the occurrence of mutator alleles (through mutation rate of an individual being modulated by a locus) and even plasticity in mutational rate (through a genotype specific variation of the mutation rate with the environment).

Computational advantages

- Mutations always happen in O(1) time with only a single conditional: it is extremely efficient irrespective of the complexity of the mutation scheme.
- Computationally, it is very fast to simulate taking time proportional to O(P+O) where P and O stand for number of parents and offspring.

Incorporation of other biological phenomena that are not obviously tied to mutation

In this model the genotype is just the collection of "properties derived from a parent that affect the phenotype" ("heritome"). Properties like epigenetic modifications, cultural inheritance and maternal effects can easily fall under this definition. As conceived in the present framework, these phenomena are identical to the DNA sequences with respect to how they are incorporated into the derivation of the final phenotypes and fitness. Where (if) they differ is at how *variation* is generated. The relative transition weights between DNA alleles remains relatively fixed, i.e despite the rate of mutation scaling, the biases towards particular alleles doesn't vary.

Epigenetic changes that are brought about by specific cues (Eva Jablonka and Raz, 2009) or maternal effects (Dantzer et al., 2013; Galloway, 2005) that depend on the parent's phenotype (e.g. food provisioned for offspring) can be described fairly completely as changes in the transition biases between the parent's and the offspring's property brought about by the state of the parent.

Instead of an Locus × Allele × New-allele-probability tensor, if the transition-weights are drawn from Genetic-background/Environmental-background × Property × Parent-State × New-state-probability tensor, this mutational scheme can trivially be expanded so the model can describe phenomena like:

- Maternal effects
- Epigenetics
- Cultural inheritance
- Niche construction

3.2.5 Summary of the most important results and future work

1. This model/framework can be used to treat many different evolutionary phenomena similarly. In brief its main uses are:
    1. The study of the evolution of G-P-F map characteristics.
    2. The unified treatment of traditional genetics, cumulative fitness, behavioural inheritance, epigenetics and plasticity.
2. Faster fluctuations in the environment may result in *reduced* selection for high mutation rates.
    1. It is proposed here that the very fast fluctuations or small time periods limit the difference in the time of the creation of new mutants when comparing different mutation rates. This reduces the payoff of higher mutation rates and relaxes selection.

2. It needs to be seen if this first order approximation is actually a significant part of the mechanism involved, and how this phenomenon relates to different selection levels and population size. Then it could be seen if predictions could be extrapolated to living populations.

4. References

Babu, M.M., Luscombe, N.M., Aravind, L., Gerstein, M., and Teichmann, S.A. (2004). Structure and evolution of transcriptional regulatory networks. Curr. Opin. Struct. Biol. *14*, 283–291.

Basler, G., Grimbs, S., Ebenhöh, O., Selbig, J., and Nikoloski, Z. (2012). Evolutionary significance of metabolic network properties. J. R. Soc. Interface *9*, 1168–1176.

Burch, C.L., and Chao, L. (1999). Evolution by Small Steps and Rugged Landscapes in the RNA Virus ϕ6. Genetics *151*, 921–927.

Crombach, A., and Hogeweg, P. (2008). Evolution of evolvability in gene regulatory networks. PLoS Comput. Biol. *4*, e1000112.

Dantzer, B., Newman, A.E.M., Boonstra, R., Palme, R., Boutin, S., Humphries, M.M., and McAdam, A.G. (2013). Density Triggers Maternal Hormones That Increase Adaptive Offspring Growth in a Wild Mammal. Science *340*, 1215–1217.

Denamur, E., and Matic, I. (2006). Evolution of mutation rates in bacteria. Mol. Microbiol. *60*, 820–827.

Ellegren, H. (2000). Microsatellite mutations in the germline:: implications for evolutionary inference. Trends Genet. *16*, 551–558.

Eva Jablonka, B., and Raz, G. (2009). Transgenerational Epigenetic Inheritance: Prevalence, Mechanisms, and Implications for the Study of Heredity and Evolution. Q. Rev. Biol. *84*, 131–176.

Falconer, D.S., and Mackay, T.F.C. (1996). Introduction to Quantitative Genetics (Longman).

Fisher, R.A. (1919). XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. Earth Environ. Sci. Trans. R. Soc. Edinb. *52*, 399–433.

G, J.A., Visser, M. de, Zeyl, C.W., Gerrish, P.J., Blanchard, J.L., and Lenski, R.E. (1999). Diminishing Returns from Mutation Supply Rate in Asexual Populations. Science *283*, 404–406.

Galloway, L.F. (2005). Maternal effects provide phenotypic adaptation to local environmental conditions. New Phytol. *166*, 93–99.

Green, P., Ewing, B., Miller, W., Thomas, P.J., NISC Comparative Sequencing Program, and Green, E.D. (2003). Transcription-associated mutational asymmetry in mammalian evolution. Nat. Genet. *33*, 514–517.

Mallarino, R., and Abzhanov, A. (2012). Paths Less Traveled: Evo-Devo Approaches to Investigating Animal Morphological Evolution. Annu. Rev. Cell Dev. Biol. *28*, 743–763.

Marais, G. (2003). Biased gene conversion: implications for genome and sex evolution. Trends Genet. *19*, 330–338.

Østman, B., and Adami, C. (2013). Predicting evolution and visualizing high-dimensional fitness landscapes. ArXiv13022906 Nlin Q-Bio.

Pigliucci, M. (2009). An Extended Synthesis for Evolutionary Biology. Ann. N. Y. Acad. Sci. *1168*, 218–228.

Pigliucci, M. (2010). Genotype–Phenotype Mapping and the End of the “genes as Blueprint” Metaphor. Philos. Trans. R. Soc. B Biol. Sci. *365*, 557–566.

Schlötterer, C., Ritter, R., Harr, B., and Brem, G. (1998). High mutation rate of a long microsatellite allele in Drosophila melanogaster provides evidence for allele-specific mutation rates. Mol. Biol. Evol. *15*, 1269–1274.

Shaver, A.C., Dombrowski, P.G., Sweeney, J.Y., Treis, T., Zappala, R.M., and Sniegowski, P.D. (2002). Fitness Evolution and the Rise of Mutator Alleles in Experimental Escherichia Coli Populations. Genetics *162*, 557–566.

Vannimwegen, E., and Crutchfield, J. (2000). Metastable Evolutionary Dynamics: Crossing Fitness Barriers or Escaping via Neutral Paths? Bull. Math. Biol. *62*, 799–848.

Vose, M.D. (1991). A linear algorithm for generating random numbers with a given distribution. IEEE Trans. Softw. Eng. *17*, 972–975.

Wielgoss, S., Barrick, J.E., Tenaillon, O., Wiser, M.J., Dittmar, W.J., Cruveiller, S., Chane-Woon-Ming, B., Medigue, C., Lenski, R.E., and Schneider, D. (2013). Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. Proc. Natl. Acad. Sci. U. S. A. *110*, 222–227.

Yu, S., Pritchard, M., Kremer, E., Lynch, M., Nancarrow, J., Baker, E., Holman, K., Mulley, J., Warren, S., Schlessinger, D., et al. (1991). Fragile X genotype characterized by an unstable region of DNA. Science *252*, 1179–1181.