# DYNAMICS IN THE MUTATION-SELECTION BALANCE IN HAPLOIDS: AN APPROACH TO THE FITNESS COMPOSITION OF INDIVIDUAL GENES IN AN INFINITE ASEXUAL POPULATION

**Anna Fukshansky**[1]

[1]Royal Holloway University of London, Computer Science Department
Egham, Surrey, TW20 0EX, UK

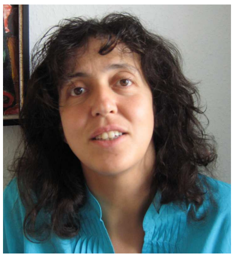
*annaf@cs.rhul.ac.uk(Anna Fukshansky)*

## Abstract

The paper presents an original approach to the whole genome treatment of an asexual haploid population under mutation and selection.
The genes are partitioned into short stretches of nucleotides, the *elements*, which have different degrees of intolerance toward mutations. In the first step of the analysis a genome is considered as a finite set of elements deprived of their association with genes. In particular, the *fitness composition of the genome*, i.e. the partition of the deleterious elements into different fitness classes, can be described in each generation and the equilibrium.
The second and third steps describe a novel effect: A displacement of the fitness composition in individual genes, which also continues in the mutation-selection balance. Together with the finiteness of the genome this allows for the retrieval of the information abandoned in the first part: sets of elements which are classified with respect to fitness can be re-associated with genome components and finally with individual genes. Hence, the most probable fitness composition of individual genes in the equilibrium can be determined.
In the second part of the present paper an algorithm is described which allows to simulate genomes using the model outlined above. The program accepts as input the description of a genome by the declaration of its genes (its length and the tolerance towards mutations of its elements). Using the model and additional methods like dynamic programming, the algorithm predicts the fitness composition of individual genes in the equilibrium. Examples show in particular, that a change in one gene has an effect on the other genes which remain unchanged.

**Keywords: population genetics, whole genome microevolution, mathematical modelling, fitness, mutation-selection equilibrium.**

## Presenting Author's Biography



Anna Fukshansky.
Born in St Petersburg, Russia.
Diploma in Mathematics, University of Freiburg, Germany 1992.
Assistant researcher at University of Halle, Germany
(finite group theory), 1995-1998).
Dr. rer. nat. in Mathematics (University of Freiburg), 1998.
Since 1998 lecturer in Computer Science
at Royal Holloway University of London.

ISBN 978-3-901608-32-2

## 1 Introduction

Classical population genetics usually considered only one or very few loci over generations in a population. However, selection acts on the whole genome and not on individual genes. To consider many genes in detail at the same time is very difficult, and mathematically only feasible for rather small numbers of genes ([1]). There are different global approaches to model the genomes of populations over generations and in the steady state (for an overview see e.g. [2]), one of the most important ones is the neutral theory created by Kimura (e.g. [3]), which, however, ignores the impact of selection, as all mutations are assumed to be neutral.

The presented model belongs to the wide class of selection-mutation models initiated by the work of Kimura and Maruyama [4] (but the idea can be traced back to Haldane [5]). These models were designed to approach different problems, among others, the evolution of muation rates ([6], [7], [8], [9], [10]). Extended by the description of a relation between fitness and the value of a multi-locus trait this type of model was applied to problems in quantitative genetics like evolution of reproduction, maintenance of phenotypic and molecular variability and evolution of mating preferences (e.g. [11]).

In the present paper a model is suggested that describes a population which is affected by mutation and selection over generations and where there are still dynamics in the steady state. These changes in the equilibrium are described and used to return from the very global and anonymous approach to the identification of individual genes and their fitness. The model allows simulations of explicit genomes made up from genes and predicts the most probable fitness composition of each gene.

## 2 The Problem and the Model

An effectively infinite strictly asexual haploid population with non-overlapping generations in a constant environment is considered. The genome of an individual is a union of $J$ genes, which is exposed to successive selection and mutation steps in each generation. The genome may also contain non-coding regions, which can be considered as a subset of genes with no function.

The genome is considered as a set of elements, each of which is a short stretch of nucleotides, chosen to be significantly smaller than a gene size and such that any element is part of only one gene.

Thus, the genome is modelled as a finite set $\mathbf{M}$ of elements, which is partitioned into the genes $\mathbf{G}_j$, $j = 1, \ldots, J$, which are constant as sets and which make a partition of the set $\mathbf{M}$:

$$\mathbf{M} = \bigcup_{j=1}^{J} \mathbf{G}_j.$$

There are $N+1$ fitness classes. In any generation, each element is assigned a local fitness from one of these classes. There is no epistasis, i.e., the fitness of a gene is the product of the fitnesses of its elements, and the fitness of a genome is the product of the fitnesses of all the elements, in particular, the fitness of the genome is the product of the fitnesses of its genes. Elements in class $n$ have local fitness $s_n$ for all $0 \leq n \leq N$. The class 0 has fitness $s_0 = 1$ whereas the other classes have fitnesses $1 > s_1 > s_2 > \cdots > s_N > 0$ and contain deleterious elements.

Mutations change the fitness of individual elements. A mutation hits only one element at a time, only point mutations are treated here. Any element mutates with the probability $u$, and the mutation may transfer the element into a different fitness class. Thus, recurrent and, hence, positive mutations are possible.

In addition there is a third partition of the elements which reflects the structural feature of a genome called the functional constraint is introduced. The elements of the genome are assigned to $I$ *risk classes* $\mathbf{R}^{(i)}, i = 1, \ldots, I$:

$$\mathbf{M} = \bigcup_{i=1}^{I} \mathbf{R}^{(i)}.$$

A risk class $\mathbf{R}^{(i)}$ is equipped with a set of parameters

$$\Gamma^{(i)} = \{\gamma_{i0}, \gamma_{i1}, \ldots, \gamma_{iN}\},$$

with $\sum_{n=0}^{N} \gamma_{in} = 1$. These parameters, called the *mutation priorities*, regulate the behaviour of an element as follows: When an element in risk class $\mathbf{R}^{(i)}$ mutates, it transfers into fitness class $n$ with probability $\gamma_{in}$.

The partition of the genome into the $I$ risk classes is fixed throughout the dynamics.

Individual genes are themselves partitioned into risk classes: For any $j = 1, \ldots, J$, the gene $\mathbf{G}_j$ is divided into fixed disjoint subsets, $\mathbf{G}_j^{(i)}$ $(i = 1, \ldots, I)$, which are called *gene risk subsets*, such that

$$\mathbf{G}_j^{(i)} = \mathbf{R}^{(i)} \cap \mathbf{G}_j, \qquad \mathbf{G}_j = \bigcup_{i=1}^{I} \mathbf{G}_j^{(i)}. \tag{1}$$

Similarly, each risk class $\mathbf{R}^{(i)}$ can be expressed as a disjoint union of the corresponding gene risk subsets $\mathbf{G}_j^{(i)}$ over all the genes, i.e.,

$$\mathbf{R}^{(i)} = \bigcup_{j=1}^{J} \mathbf{G}_j^{(i)}. \tag{2}$$

So far the necessary notation and the building blocks of the model are established. It should be pointed out that the finiteness of the genome is essential for the next section, where the three steps are described which lead to the fitness composition of individual genes.

## 3 Fitness composition of individual genes

It is possible to derive the *fitness composition* acquired in the mutation-selection balance by each gene, i.e., the distribution of the elements into fitness classes for each gene. Assuming that a partition of the genome into risk classes is fixed, this distribution is determined in three major steps.

1. In the first step the functionality of the elements is ignored and the genome is considered as a set of anonymous elements, whose dynamics are described. These results are adapted from [12] and yield a description of the equilibrium, in particular the distribution of the number of deleterious elements is known, as well as the expected number of elements, $A_n$, in each fitness class $n$.

2. In the second step, the fitness composition of the individual risk classes in the mutation-selection balance is retrieved. This retrieval procedure is mainly possible, because the genome in the model is finite. In the equilibrium two partitions of the elements of the genome are considered: On the one hand, the partitioning into risk classes; on the other hand from the former step the expected numbers of elements in each fitness class are known. Let the number of elements in risk class $\mathbf{R}^{(i)}$ which belong to fitness class $n$ be denoted by $y_{in}$. In the equilibrium mutation and selection still affect the population, and the parameters $\{y_{in}\}$ are changing, but such that the following holds

$$\begin{aligned} &\sum_{n=0}^{N} y_{in} = |\mathbf{R}^{(i)}|,\ (i = 1, \ldots, I) \quad \text{and} \\ &\sum_{i=1}^{I} y_{in} = A_n,\ (n = 0, \ldots, N). \end{aligned} \tag{3}$$

   It turns out that there is only one kind of change which obeys the following three conditions:

   (a) The left side of (3), which holds per definition of risk classes with a fixed number of elements, is satisfied;

   (b) A change corresponds to the physical nature of mutation, and

   (c) Only the changes surviving the balancing selection are considered; this means, the right side of (3) is satisfied.

   These changes to the set $\{y_{in}\}$ are analysed.

   A discrete time scale is introduced, such that at any time (3) is true and that during each time step at most two independent transfers (due to mutations) can take place. The situation can be described by the following double constrained random process visualised by the graph in figure 1.

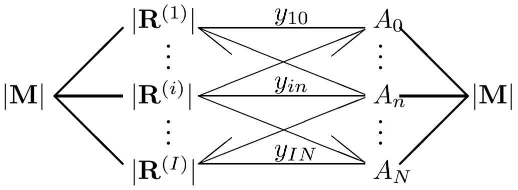


Fig. 1 *The elements of* $\mathbf{M}$ *are partitioned into risk classes* $\mathbf{R}^{(i)}$ *(on the left) and into fitness classes (on the right). The number of elements in* $\mathbf{R}^{(i)}$ *and in fitness class* $n$ *is* $y_{in}$.

   The most probable set $\{y_{in}\}$ is determined applying the balancing method, the convergence of which was proved in [13] (this method is also used e.g. in [14], [15]).

   It turns out that this most probable set $\{y_{in}\}$ coincides with the minimum of the weighted Kullback entropy.

3. In the third step the connection between the elements and their genes is recovered and the fitness composition of individual genes is derived. Here a similar idea as in the previous step is used. This time two partitions of the elements in an individual risk class $\mathbf{R}^{(i)}$ are considered – on the one hand the distribution into gene risk subsets (see equation (2)), and on the other hand the most probable solution of the previous step: the expected number of elements $y_{in}$ in each fitness class in risk class $\mathbf{R}^{(i)}$. The most probable fitness composition of all gene risk subsets is determined.

   As in the previous step, this solution is independent of the initial state and allows to look for the most probable composition $\{g^*_{ijn}\}$. The most probable solution this time turns out to be a shifted uniform distribution, where the shift is given by the maximum of the Shannon measure under the implicit constraints.

   From here, the fitness composition of each gene, which is a disjoint union of its gene risk subsets (see (1)) is derived.

These three steps present a model which is a theoretical tool that enables to find the fitness compositions of individual genes. In the following section a brief description of a program which allows to experiment using this theory.

## 4 Simulation of genomes in the equilibrium

The experimental part of this work consists of simulations of different compositions of genomes in the equilibrium and the calculations of the fitness compositions of individual genes. The simulation is based on the theoretical model, the idea of which is described in the former sections.

### 4.1 Some Notation

Let $N$ be the number of deleterious fitness classes for elements. The local fitness in class $n$ is

$$s_n = e^{\Delta n}, \quad n = 0, 1, \ldots, N;$$

where $\Delta$ or more intuitively, $e^{-\Delta}$ is a parameter that can be chosen later. The genome consists of $|\mathbf{M}|$ elements each of which can be in one of the $N+1$ fitness classes. The elements contribute multiplicatively to the fitness class of the genome. Since

$$s_{n_1} * s_{n_2} = e^{\Delta n_1 + \Delta n_2} = e^{\Delta(n_1+n_2)},$$

the fitness of the whole genome can only be one of the following:

$$s_n = e^{\Delta n}, \quad n = 0, 1, \ldots, N * |\mathbf{M}|;$$

in particular, the worst fitness for the genome is $s_{N*|\mathbf{M}|}$, this is the case, where all the $|\mathbf{M}|$ elements lie in the worst fitness class $N$.

A genome is represented by a vector $\vec{y} = (y_1, \ldots, y_N)$, where $y_n$ is the number of deleterious elements which have fitness $s_n$, i.e., the number of elements, which are in fitness class $n$. Given such a vector, the fitness of the genome is immediately found:

$$s_{genome} = e^{1\Delta \cdot y_1} e^{2\Delta \cdot y_2} \cdots e^{N\Delta \cdot y_N} = \\ = e^{\Delta(y_1+2y_2+\cdots+Ny_N)} = s_{(y_1+2y_2+\cdots+Ny_N)}.$$

It is necessary for further analysis to find the probability for the whole genome and for individual genes to be in a certain fitness class. So, as the first step towards this goal for each number (potential fitness class of the genome or gene) all the possible sum representations of the type $(y_1 + 2y_2 + \cdots + Ny_N)$ have to be determined. The algorithm which finds all these representations for a part of or the whole genome is described in the next section.

### 4.2 Determine for all numbers the possible representations as sums

In this section the recursive algorithm to find all the possible representations of any number $n > N$ in the form $y_1 + 2y_2 + \cdots + Ny_N$ is described.
Let the set of the vectors $(y_1, y_2, \ldots, y_N)$ which satisfy $y_1 + 2y_2 + \cdots + Ny_N = n$ be denoted by $\Theta_n$. The main observation which is necessary for the recursion is the following: The set $\Theta_n$ can be determined if all the representations $\Theta_{n-1}, \Theta_{n-2}, \ldots, \Theta_{n-N}$ are known. In fact, for large enough $n$, the set $\Theta_n$ of representations is constructed as follows:

1. Initially $\Theta_n$ is empty.
2. From the set $\Theta_{n-1}$, all the vectors are chosen, the number $y_1$ is substituted by $y_1 + 1$, and the resulting vectors are added to $\Theta_n$.
3. From the set $\Theta_{n-2}$, those vectors which have $y_1 = 0$ are chosen, $y_2$ is increased by one, and the vectors are added to $\Theta_n$.

   Note, that only those vectors with $y_1 = 0$ have to be taken into account, as those with $y_1 \neq 0$ are already present from the set $\Theta_{n-1}$.
4. From the set $\Theta_{n-k}$ for $2 < k \leq N$, those vectors which have $y_1, \ldots, y_{k-1} = 0$ are chosen, $y_k$ is increased by one, and the vectors are added to $\Theta_n$.

Using this recursion, a dynamic programming algorithm is easily derived. The first sets $\Theta_1, \ldots, \Theta_N$ are constructed by hand. Then, recursively, the sets for $N+1, N+2, \ldots,$ in this order are constructed, using the recursion above. In that way, the necessary $N$ predecessor sets are always present for each new construction.

Now for any genome fitness class, all the possible vectors $(y_1, y_2, \ldots, y_N)$ are known, in other words, for any desired fitness for a genome, all possible combinations of numbers of elements in fitness classes are known. Note, that the number of elements in the best fitness class 0 is given by $|\mathbf{M}| - \sum_n y_n$.

In the next section the sets $\Theta_n$ are used to determine $P(s_n)$, the probability for a genome to have fitness $s_n$ for all $n \leq |\mathbf{M}| * N$.

### 4.3 Determine the probability of a gene to have some given fitness

In this section the concluding step of the algorithm is shown: The probability for a genome to have fitness $s_n$, this is $P(s_n)$ in the present notation, is determined. This is the sum of the probabilities of all the possible combinations $(y_1, y_2, \ldots, y_N)$ in this fitness class.

The probability that a genome is given by a vector $(y_1, y_2, \ldots, y_N)$ is equal to

$$P(\vec{y}) = \frac{(\sum_n y_n)! \prod_n \alpha_n^{y_n}}{\prod_n y_n!}.$$

Hence, the probability $P(s_n)$ is given by

$$P(s_n) = \sum_{\vec{y} \in \Theta_n} \frac{(\sum_n y_n)! \prod_n \alpha_n^{y_n}}{\prod_n y_n!} \qquad (4)$$

Since the set $\Theta_n$ is known (section 4.2), the probability $P(s_n)$ is determined in a loop over the elements fo $\Theta_n$.

So far, the whole genome and its probability to belong to a certain fitness class was considered. Now a similar treatment is done for individual genes. Note, that the sum representation (section 3) is used for genes as well as for the whole genome.

### 4.4 Fitness distribution for individual genes

In the previous section the probability was found for a genome to lie in a certain fitness class. In equation (4) the steady state parameters $\alpha_n$ are necessary for this. For individual genes the expected number of elements in different fitness classes is not part of the description in the steady state (after the anonymous fitness dynamics), because the genes are not considered yet. However, after the two steps of post-selectional dynamics the limit fitness composition is given as follows: The distribution, $P_j^{(i)}(X_0, X_1, \ldots, X_N)$ of the subset $\mathbf{G}_j^{(i)}$ into fitness classes is a polynomial distribution with the probability parameters

$$p_{ijn} = \frac{g^*_{ijn}}{|\mathbf{G}_j^{(i)}|}, \quad n = 0, \ldots, N.$$

A gene is the disjoint union of its risk gene subsets, hence, the distribution of the gene $\mathbf{G}_j$ is also polynomial, $P_j(X_0, X_1, X_2, \ldots, X_N)$, and the parameters are

$$p_{jn} = (\sum_i g^*_{ijn})/|Gj|.$$

These parameters play the same role for the gene as the $\alpha_n$ for the whole genome. However, there is a slight difference, as for the genome, the $\alpha_n$ are for $n = 1, 2, \ldots, N$, while for risk gene subsets and genes the fitness class $n = 0$ is also included.

For any vector $\vec{y} = (y_0, y_1, y_2, \ldots, y_N)$ the probability is given by

$$P(\vec{y}) = \frac{|\mathbf{G}_j|! \prod_n p_{jn}^{y_n}}{\prod_n y_n!}.$$

The probability for a gene to belong to fitness class $n$ is

$$P(s_n) = \sum_{\vec{y} \in \Theta_n} \frac{|\mathbf{G}_j|! \prod_n p_{jn}^{y_n}}{\prod_n y_n!}. \qquad (5)$$

This is calculated using the sum presentation from section 4.2 for each potential fitness class $n$.

The probabilities $P(s_n)$ for large $n$ are very small. Hence, as a timesaving approximation the highest number $n^*$ is determined such that the sum of all $P(s_n)$ for $n = 0, 1, \ldots n^*$ is $> 0.9999$. The $P(s_n)$ is set to $0$ for $n > n*$.

### 4.5 Summary and Input/Output of the fitness composition program

So far it is possible for example to submit a genome with a risk structure, and genes to the program, and the result will be the most probable fitness composition of each gene. In this section the description of the necessary input parameters and possible output values is given.

The global necessary input parameters are:

1. Probability for an element to mutate, $u$.
2. Number of Genes, $J$ and their lengths (how many elements does each gene have). In particular, this yields the number of elements $|\mathbf{M}|$ in the genome, this is the sum of the lengths of all genes.
3. Number of risk classes, $I$.
4. Number of element fitness classes, $N$.
5. Fitness step $e^{-\Delta}$, the fitness in fitness class 1, the local fitnesses in the other classes are determined from it ($e^{-n\Delta}$ in class $n$).

In addition, the risk structure has to be chosen. This is done by

6. Defining each risk class $\mathbf{R}^{(i)}$ by the associated vector $\Gamma_i = (\gamma_{i0}, \ldots, \gamma_{in}, \ldots, \gamma_{iN})$.
7. Determine for each gene, how many elements lie in which risk class, i.e., the size of the risk gene subsets $\mathbf{G}_j^{(i)}$ for all $i, j$ has to be entered.

The program determines according to the formulae in the steady state after the anonymous dynamics the following parameters:

1. The expected vector $(\alpha_1, \ldots, \alpha_N)$;
2. The probability distribution of the number of deleterious elements, in particular its mean $\mu$.
3. From there the expected sizes of the fitness classes, $A_n$ for $n = 0, \ldots, N$.
4. Using the balancing method twice, and using the $A_n$ the program calculates the most probable fitness composition for each gene risk subset $\mathbf{G}_j^{(i)}$, which is given by the parameters $g^*_{ijn}$.
5. Using the fitness composition of gene risk subsets, the expected numbers of elements in fitness classes, $p_{jn}, n = 0, 1, \ldots, N$ in individual genes is determined.
6. With the parameters $p_{jn}$ for all genes and using the sets $\Theta_n$, as described in section 3, for each gene a file containing the probabilities of all $P(s_n)$ is constructed. The highest $n$ is the last one, which is necessary, such that the sum of all $P(s_n)$ is $\geq$ 0.9999.

When running the simulations different aspects can be observed. For example, and not surprisingly, the increase of the mutation rate $u$. reduces the local fitness of the genes in the steady state.

There are different possibilities to observe A change of the risk structure of a single gene and its influence on the steady state fitness of the other unchanged genes. The effect depends on the choice of the risk structure before and after the change in the chosen gene as follows. At first the gene consisted of elements, most of which were in a good risk class, i.e., the probability to mutate into a good fitness class was high. With a change of the majority of its elements into a bad risk class (most mutations lead the gene's elements into bad fitness classes), the gene's fitness obviously was reduced, but the fitness of all the other genes was increased.

## 5 Outlook

In this paper the theory that allows to find fitness compositions of individual genes is outlined. The major novelty in this approach is the global view, that considers a genome as a set of finitely many elements. At first the connection between elements and genes is ignored, only in the equilibrium it is re-established and leads to the expected fitness composition of individual genes.
The simulation part of the work allows to observe in particular the influence of changes in individual genes on the other genes in the genome. For now these changes are done only in terms of the local fitness contribution of the genes. However, the next level of identifying genes using this theory is to introduce alleles and to find allele frequencies of individual genes using their fitness compositions. With this extended model, it will also be possible to observe not only the changes in fitness but also in allele frequencies. This extended model is work in progress of the author at the moment.

## 6 References

[1] N.H. Barton and M. Shpak. The stability of symmetric solutions to polygenic models. *Theoretical Population Biology*, 57:249–263, 2000.

[2] R. Bürger. *The mathematical theory of selection, recombination, and mutation*. John Wiley & Sons, New York, 2001.

[3] M. Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge, University Press, Cambridge, 1983.

[4] M. Kimuara and T. Maruyama. The mutational load with epistatic gene interactions in fitness. *Genetics*, 54:1337–1351, 1966.

[5] J. B. S. Haldane. The effect of variation on fitness. *Am. Naturalist*, 71:337–349, 1937.

[6] A. S. Kondrashov. Modifiers of mutation-selection balance: General approach and the evolution of mutation rates. *Genet. Res.*, 66:53–77, 1995.

[7] K.J. Dawson. Evolutionarily stable mutation rates. *J. Theor. Biol.*, 194:143–157, 1998.

[8] K.J. Dawson. The dynamics of infinitesimally rare alleles, applied to the evolution of mutation rates and the expression of deleterious mutations. *Theor. Population Biology*, 55:1–22, 1999.

[9] T. Johnson. Beneficial mutations, hitchhiking and the evolution of mutation rates in sexual populations. *Genetics*, 151:1621–1631, 1999.

[10] T. Johnson. The approach to mutation-selection balance in an infinite asexual population, and the evolution of mutation rates. *Proc. R. Soc. Lond.*, B 266:2389–2397, 1999.

[11] A. S. Kondrashov. Deleterious mutations and the evolution of sexual reproduction. *Nature*, 336:435–440, 1988.

[12] A. Fukshansky. A stochastic model for the mutation-selection balance in an infinite asexual population with a genome of fixed size. *Journal of Theor. Biology*, 231, 4:557–562, 2004.

[13] L.M. Bregman. Proof of the convergence of sheleikhovskii's method for a problem with transportation constraints. *USSR Computational Mathematics and Mathematical Physics*, 7(1):191–204, 1967.

[14] B.G. Pittel. A simple probability model of collective behaviour. *USSR Problems of Information Transmission (Problemy peredachi informatsii)*, 3(3):37–52, 1967.

[15] L. Fukshansky. A stochastic approach to multicomponent lipid-protein interactions. In *on New Developments and Methods in Membrane Research and Biol. Energy Transduction*, pages 1358–1359, Island of Spetsai, Greece, August 16-29 1984. Proc. NATO Summer School.