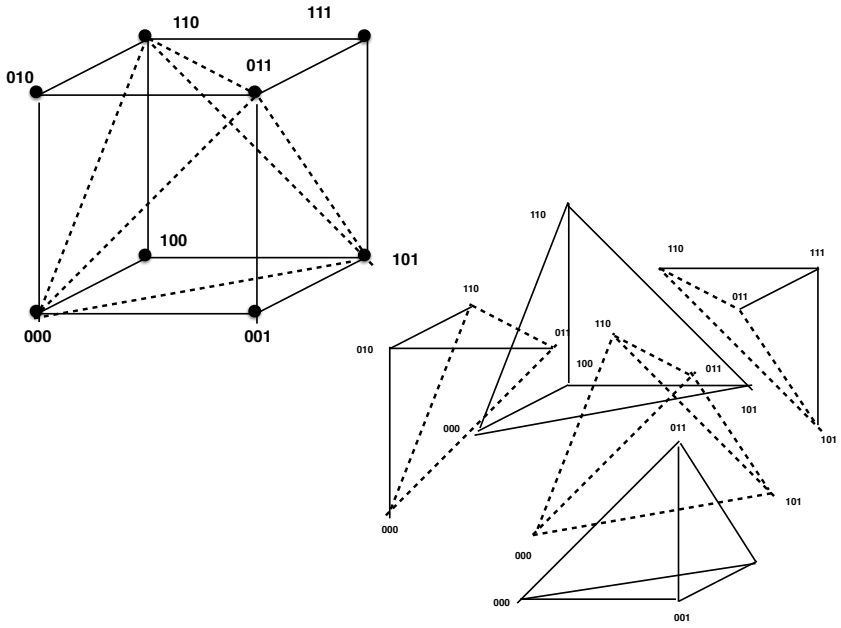
EPISTASIS, SHAPES & EVOLUTION

MASTER'S THESIS



MALVIKA SRIVASTAVA

Institute for Biological Physics
University of Cologne

DECEMBER 2018

Supervisor:
Prof. Dr. Joachim Krug

Second Corrector:
Prof. Dr. Thomas Wiehe

The figure on the cover page depicts type 1 triangulation of the three locus Genotope. The fitness landscape that imposes this triangulation has shape 2.

Abstract

Along with being consequential for evolution, epistasis is also quite prevalent in nature. Thus, it is important to study it. Till date, there exist many methods of inferring epistasis from experimental and theoretical fitness landscapes. The theory of *shapes* of fitness landscapes is another addition to that list. In this thesis, the shape theory of fitness landscapes is first introduced and then compared to pre-existing methods of gauging epistasis. From such a comparison for 3 locus landscapes, it turns out that landscapes of different interaction *types* differ in ruggedness, number of reciprocal sign epistasis motifs and presence of higher order epistasis. Next, the applicability of shapes in studying empirical fitness landscapes is explored. Here the theory proves to be useful because the additional tests suggested by the Markov basis further corroborate the diminishing returns epistasis hypothesis, especially for the β-lactamase landscape with synonymous mutations. Moreover, the triangulation of the landscape of large effect mutations has a particular genotype as vertex of every tetrahedra in the triangulation, indicating the presence of that genotype in all fittest populations. Finally, the effect of the shape on the evolutionary dynamics is discussed. For two locus landscapes, the equilibration time of the mutation-selection dynamics has a sharpness exactly at the transition point between the two shapes. Further, it was found that Eshel and Feldman's results regarding the advantage of recombination in two locus permutation invariant landscapes can be extended to three locus landscapes. It turns out that in three out of the six shapes of permutation invariant landscapes, recombination is "advantageous", while in the other three, it is "disadvantageous". This extensive analysis of its applicability indicates that the shape theory offers useful insights while studying empirical landscapes, however additional constraints are needed to predict evolution on landscapes of different shapes.

Erklärung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Alle Stellen, die wörtlich oder sinngemäßaus veröffentlichten und nicht veröffentlichten Schriften entnommen wurden, sind als solche kenntlich gemacht. Die Arbeit ist in gleicher oder ähnlicher Form oder auszugsweise im Rahmen einer anderen Prüfung noch nicht vorgelegt worden. Ich versichere, dass die eingereichte elektronische Fassung der eingereichten Druckfassung vollständig entspricht.

Köln, April 2, 2019

Malvika Srivastava

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Prof. Dr. Joachim Krug, for indicating the starting point of my thesis and then giving me the rare freedom and the accompanying opportunity to find my own path, for always giving useful inputs and for being extremely thoughtful and supportive throughout.

Next, I would like to thank all my group members and colleagues here– Alex, Benjamin, David, Jonas, Lucy and Suman, for always being there to listen, talk and help. Lucy deserves a special mention for never failing to lift my spirits when they were flagging. I want to additionally thank Benjamin for always agreeing to debug Julia and translate German documents. I would also like to thank Lara Bössinger from the mathematics department, for I learnt more from her in 2 hrs than I learnt from textbooks in 2 months. And finally Nikhil, for always pushing me to improve and develop confidence in myself, by being my best friend and my harshest critic.

Last, but definitely not the least, I want to thank my family and friends in India and elsewhere, who were yet just a phone call away. Especially, my parents for their unwavering support and love– I can never thank them enough for that, and my little sister Manya, who's now old enough to help me understand concepts from group theory. They make it all worthwhile.

Contents

Chapter 1

Introduction

By attributing our existence to accident, and not design, evolution, both removes and adds meaning to life. On one hand, it strips the human race of its narcissism, while on the other, it makes life valuable, by highlighting its sheer improbability. Also, by offering explanantions for a variety of natural phenomena, evolution makes the world we see around us, a little less surprising. So in short, evolution is the best answer we have to some of the most profound questions we have about ourselves and our surroundings. Moreover, evolution not only lends us perspective on our lives, it also serves as a guide to solving complex problems– as was testified by the recent Nobel prize in chemistry, that was awarded for using directed evolution in the lab to develop useful chemicals [1].

1.1 Forces of evolution

In a nutshell, evolution is driven by the forces of selection, mutation, recombination, genetic drift and migration. While mutation, recombination and migration are responsible for introducing diversity on which selection can act, genetic drift accounts for the inevitable fluctuations in the dynamics of finite populations. In this thesis, only the forces of selection, mutation and recombination are included and the population size is considered to be infinite, which means that the population dynamics is deterministic.

1.2 Fitness landscapes

Evolutionary processes span many length and time scales. Even events occurring on small scales, for instance mutation in a single base pair, can have

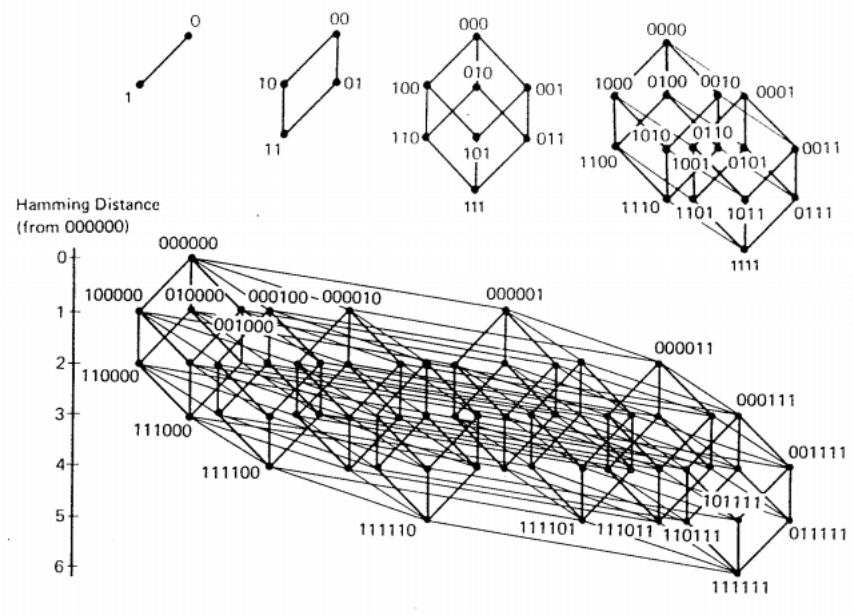


Figure 1.1: The one, two, three, four and six dimensional Hamming spaces. Source: [2]

cascading effects on the overall *fitness* of an organism [3]. So a natural question that arises is how can we study this multi-scale process? This is where the concept of *fitness landscapes* comes into the picture.

The term fitness can mean different things in different contexts [4]. While it typically refers to the fecundity of an organism, it could also refer to the viability in studies of age structured populations or the minimum inhibitory concentration (MIC) in studies of antibiotic resistance. Regardless of its definition, fitness of an organism is co-determined by its DNA sequence or its *genotype* and the environment in which it evolves. This means that for a constant environment, there exists a map from the genotype of an organism to its fitness.

The genotype of a haploid organism can be simply modelled as a sequence of a fixed number of sites L, with a fixed number of alleles a, at each site. Each site can represent, for example, a nucleotide (i.e. A,G,T or C $\Rightarrow a = 4$) or even an entire gene. The *genotype space* G, is then the set of all possible sequences of length L that can be formed by combining the a alleles at each site and $|G| = a^L$. Further, a metric can be defined for the genotype space in terms of the number of sites at which two sequences differ. In the case of a alleles, the

genotype space can be mapped to the L dimensional, a-allelic Hamming space ($\mathbb{H}_a^L$) [5] and the metric then is the Hamming distance

$$d(\sigma,\gamma) = \sum_{i=1}^{L}(1-\delta_{\sigma_i\gamma_i}) \tag{1.1}$$

where σ and γ are sequences of length L , δ_{ij} is the Kronecker delta function and σ_i, γ_i are the alleles at the ith sites on the sequences σ and γ respectively.

In this thesis, the discussion will be confined to bi-allelic sequences, where the L dimensional binary Hamming space $\mathbb{H}_2^L$ can be represented by the vertices of the L-dimensional hypercube (figure 1.1).

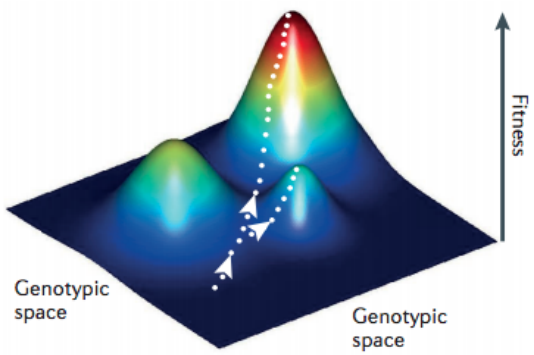


Figure 1.2: An illustration of a fitness landscape. The white arrows represent evolutionary trajectories. Source:[6]

With that we have all three "ingredients" [7] required to define a fitness landscape, namely, a configuration space i.e. G, a notion of distance between the elements of the space i.e. $d(\sigma,\gamma)$ and a map from every element σ of G to the fitness, $F(\sigma) \in \mathbb{R}$. A fitness landscape is then defined as $F : G \to \mathbb{R}$.

Figure 1.2 shows a fitness landscape, like it was first imagined by Sewall Wright [8]. Wright himself realized that it is an inadequate representation of the true higher dimensional picture because it constrains the genotype space to be only two dimensional.

Theroetical fitness landscapes can be modelled in several ways [6]. Following are two of the most commonly used models:

- **HoC model:** The simplest model is called the House of Cards (HoC) model [9] and it assumes every fitness value to be an independent identically distributed (i.i.d) random variable.

- **NK model:** The NK model generates landscapes with tunable ruggedness and was first introduced in [10]. Here, N stands for the number of loci in the sequence (=L) and each locus interacts with K-1 other loci. Within each set of the K interacting loci, fitness contributions are assigned at random to the 2^K possibilities. The HoC model is a limiting case of the NK model when K=N.

1.3 Epistasis

Epistasis, to quote Weinreich et al. [11], is the "surprise at the phenotype when mutations are combined, given the constituent mutations' individual effects". This just means that mutations don't have independent effects. Rather, their effects depend upon the background sequence on which they occur. This makes epistasis highly consequential for evolution. In fact, many studies have already recognised the importance of epistasis for adaption, evolutionary predictability and the evolution of sexual reproduction [12, 13, 6]. Epistasis has also been linked to the topography of fitness landscapes [14, 15], which determines the accessibility of adaptive walks in sequence space. Moreover, epistasis is highly prevalent in nature and empirical fitness landscapes are known to be topographically complex [6]. This makes its inclusion in theoretical studies necessary. Lastly, the recent hypothesis that complex traits may be omnigenic [16] and that the effect of "core" genes also depends upon all the "peripheral" genes, just highlights the presence of epistasis between these genes.

1.3.1 Causes of epistasis

In [17], possible proximate and evolutionary causes of epistasis are discussed. In the past, people have used metabolic models and the concept of pleiotropy and robustness to predict and explain epistasis. In theoretical studies, epistasis has also been thought of as a dynamic variable that is subjected to evolutionary forces of selection, drift, mutation and recombination. The motivation for that is most probably Malmberg's [18] experimental system where recombination alleviated epistasis between beneficial alleles. However, despite many forward steps, the origin and dynamics of epistasis are still enigmatic.

1.3.2 Measures of epistasis

Unidimensional epistasis

One can either study unidimensional epistasis or multidimensional epistasis [19]. The unidimensional study entails looking at the mean log fitness as a function of the number of mutations. Deviations from linearity is then interpreted as epistasis. Due to the ease of measurements, most experimental studies use the unidimensional definition of epistasis. However, it fails to provide a complete picture of the underlying interactions because despite the presence of interacting loci, unidimensional epistasis can be zero. This is where multidimensional epistasis comes to use.

Multidimensional epistasis

By considering interactions between all possible combinations of loci, multidimensional epistasis provides crucial information about the number of peaks and the accessibility of fitness landscapes. It can be classified into 2 types:

1. **Pairwise epistasis**, as the name suggests, refers to epistasis between loci pairs. It can further be classified into a) Magnitude and b) Sign epistasis, based on whether the effect of a mutation has a different magnitude on a different background or whether it has an altogether different sign. A special case of sign epistasis is called reciprocal sign epistasis, wherein the sign of the mutational effect of either mutation changes in the presence of the other.

2. **Higher order epistasis** refers to interactions between more than 2 loci and it has been relatively less studied [11]. It essentially means that only knowledge of the fitnesses of the wild type, the single mutants and the double mutants is not enough to determine the rest of the fitness landscape. Like pairwise epistasis, higher order epistasis can also be classified into sign and magnitude epistasis.

   In order to assess its ubiquity, Weinreich et al. [11] analysed 14 published fitness landscapes and found that in nearly every case, the mean magnitude of higher order contributions were larger than or equal to the pairwise effects, implying that higher order epistasis is quite prevalent in nature. More recently, abundance of higher order epistasis was also found in [20] and its indispensibility in determining evolutionary trajectories was identified in [21, 22]. However, despite that, there is no

unique way of extracting information or classifying landscapes based on these interactions.

Since fitness landscapes with multiple loci have complex high dimensional structures, it is important to be able to characterize them based on simpler and preferably scalar measures. The following ways to study and classify higher order epistasis exist in the literature:

- For combinatorially complete fitness landscapes, the *Walsh coefficients* [23] can be obtained by a linear transformation of a vector containing the fitness values of all the genotypes. The first order Walsh coefficients represent the individual mutational effects averaged over all possible backgrounds, the second order coefficients represent pairwise epistasis averaged over all backgrounds and the higher order coefficients have similar interpretations.
- In [24] and [22], Crona et al. showed that higher order epistasis can also be inferred from fitness graphs which are basically directed acyclic hypercube graphs. What makes this interesting is that their analysis requires only the partial order of fitness values and not the actual values themselves. From fitness graphs, one can also extract indirect measures of epistasis, such as the number of peaks, the fraction of sign/reciprocal sign epistasis motifs etc.
- Another measure based on the correlation of mutational effects was developed in [25]. They defined an epistasis measure γ = Cor(s(g),s(g_1)), where s($g_{[i]}$) is the fitness effect of a mutation occurring at site i on the genotype g and g_1 represents neighbouring genotypes of the genotype g. Like fitness graphs, this method can also be employed to incomplete fitness landscapes, although the error in the estimate of γ increases with the fraction of unknown fitness values. But unlike fitness graphs, it can also be used to infer magnitude epistasis. Further, it proves to be different from the non-linear part of the Walsh spectrum [14] because it gives more weight to higher order epistasis than pairwise epistasis.
- Last, but hopefully not the least, is the shape theory of fitness landscapes [26]. It is also the first study that considered higher order epistasis to be important. Herein, the authors identified pairwise and higher order epistasis tests (i.e. *circuits* and *Markov bases*) that should be relevant for classifying fitness landscapes based on the kind of epistatic interactions that they exhibit. Studying the usefulness of these epistasis tests and the classification prescribed

by the authors, in comparison to the other measures, comprises one of the main motives of this thesis.

1.4 Overview of the thesis

To summarize, epistasis strongly affects both the static properties of fitness landscapes, like its ruggedness, and the dynamic properties of populations evolving on these landscapes. Although the fitness landscape is a coarse grained concept, that glosses over several intermediate levels, a lot can still be learnt from it because it's possible to extract information about mutational interactions from it. However, for multi-loci ($L > 2$) landscapes, there is no unique way of doing this. Furthermore, it is also of interest to be able to classify landscapes based on these interactions. The hope that landscapes with similar interactions will show similar static properties and population dynamics is implicit in the attempts to classify landscapes. This very hope will drive the discussion in the following chapters and the primary focus will be on the recently developed shape theory of fitness landscapes.

The organisation of the next chapters is as follows: In chapter 2, the geometric theory of fitness landscapes is introduced. Then in chapter 3, shapes are compared to other ways of classifying epistatic fitness landscapes. In chapter 4, the shape theory is applied to some empirical fitness landscapes, in order to see if some additional insights are gained from doing so. In chapter 5, the focus is on mutation-selection dynamics of populations on landscapes with different shapes. Next, the question of evolution of recombination, in the context of the shape theory, is addressed in chapter 6. Finally, in chapter 7, conclusions and the future directions are discussed.

Chapter 2

The shape theory

The shape theory of fitness landscapes was developed in [26]. The motivation of the authors was to highlight the underlying combinatorial geometry of fitness landscapes. The basic idea is that epistatic interactions between multiple loci can take place in a finite number of ways. The regular triangulations of the Genotope, encode these finite possibilities of interaction. Therefore, to quote the authors, "The biological problem of studying the genotype interactions for a fitness landscape is thus equal to the combinatorial problem of finding the shape of the fitness landscape...". However, to be able to fully understand and appreciate this statement, some mathematical foundation is built in the first section.

2.1 Mathematical preliminaries

In the following: For n points $v_1, v_2, ..., v_n$ in $\mathbb{R}^d$ $\mathrm{A} := \begin{bmatrix} v_1 & v_2 & ... & v_n \end{bmatrix} \in \mathbb{R}^{d \times n}$

Definition 2.1.1. An affine space is $\{x \in \mathbb{R}^n : B \cdot x = b\}$ where B is a $m \times n$ matrix and $b \in \mathbb{R}^m$.

Definition 2.1.2. An affine combination of a set of points $\{v_i\}$ equals $\sum \lambda_i \cdot v_i$ where $\sum \lambda_i = 1$ and $\lambda_i \in \mathbb{R}$ $\forall i$

Definition 2.1.3. A set of points is said to be affinely independent if no point in the set can be expressed as an affine combination of all the other points in the set. Else the points are affinely dependent.

Definition 2.1.4. A convex combination of a set of points is an affine combination with $\lambda_i \geq 0$ $\forall i$.

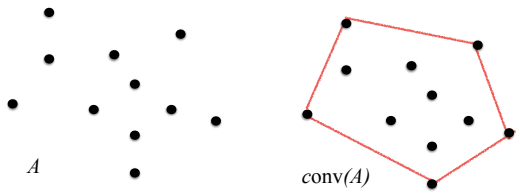


Figure 2.1: The convex hull of a given point configuration.

Definition 2.1.5. The **convex hull** of a set of points A is the set of all convex combinations of the points. It is denoted as conv(A) and is illustrated in figure 2.1.

Definition 2.1.6. There are two equivalent[1] ways of defining a **polytope**:

1. A **(V-) polytope** is the convex hull of a finite set of points.
2. An **(H-) polytope** is the intersection of half spaces[2] that must be bounded.

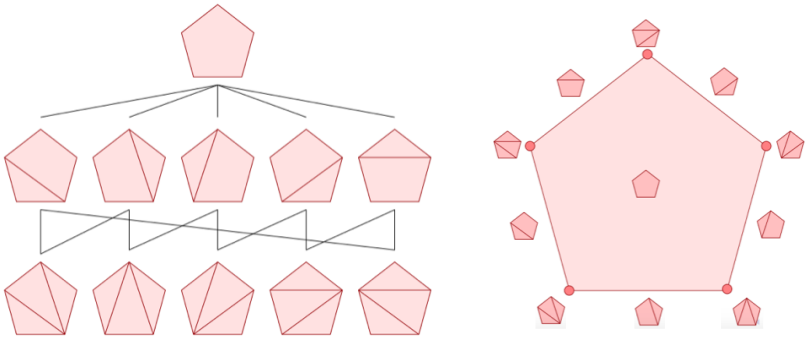
Definition 2.1.7. An n-simplex is the convex hull of $n+1$ affinely independent points, e.g. a 0-simplex is a point, 1-simplex is a line, 2-simplex is a triangle and 3-simplex is a tetrahedron.

Definition 2.1.8. A **polyhedral subdivision** of a point configuration A is a set of polytopes C such that:

1. If $c \in C$, each face of c belongs to C (closure property)
2. The union of c is equal to conv(A) (union property)
3. For $c, c' \in C$ and $c \neq c'$, the intersection of c and c' doesn't contain any interior points of c or c'. (intersection property)

Definition 2.1.9. A polyhedral subdivision is a **triangulation** if all the polytopes in C are simplices.

[1] Main Theorem of Polytope Theory [27]
[2] Any hyperplane $\vec{a} \cdot \vec{x} = b$ in $\mathbb{R}^d$ defines two half-spaces $\vec{a} \cdot \vec{x} \leq b$ and $\vec{a} \cdot \vec{x} \geq b$



(a) The poset of subdivisions (b) The secondary polytope

Figure 2.2: Subdivisions of a pentagon.

Definition 2.1.10. A **regular triangulation** is one that can be induced by a lifting construction (see figure 2.3), which in our case is a fitness landscape.

Definition 2.1.11. A **GKZ** (Gelfand, Kapranov and Zelevinsky) **vector** of a triangulation Δ of A is the vector:

$$\phi_\Delta := \sum_{i=1}^{n} (\text{vol}(\tau) : \tau \in \Delta \text{ and } i \in \tau)\vec{e}_i \in \mathbb{R}^n \tag{2.1}$$

where, vol(τ) is the normalised volume of the simplex τ, i.e. the absolute value of the determinant of A_τ divided by the greatest common divisor (g.c.d.) of the maximal minors of A.

Definition 2.1.12. A **secondary polytope** of a given point configuration A is the convex hull of the GKZ vectors corresponding to the triangulations of A. In simpler terms, a secondary polytope is a polytope whose vertices are in bijection with regular triangulations of A (see figure 2.2 b)).

Actually, the poset[3] of regular polyhedral subdivisions of a point set A equals the face poset[4] of the secondary polytope of A. Thus, triangulations are minimal elements in the poset of subdivisions.

Definition 2.1.13. If a subdivision is only refined[5] by triangulations then it is refined by exactly two of them. These two triangulations are then said to differ by a **bistellar flip**.

[3] with partial ordering induced by refinement

[4] This is the set of all faces of a polytope ordered by set inclusion.

[5] If sets $S = \{S_1 ... S_l\}$ and $T = \{T_1 ... T_m\}$ are two subdivisions of conv(A), then T is a refinement of S if $\forall j, 1 \le j \le m, \exists i, 1 \le i \le l$ such that $T_j \subseteq S_i$

These flips constitute the next to minimal elements in the poset of polyhedral subdivisions of A [28]. The poset of subdivisions of a pentagon is shown in figure 2.2 a). Essentially, flips are the minimal possible changes in the shape and are detected by minimal affine dependences in the point configuration. An interesting result is that the graph of triangulations of n points in convex position in $\mathbb{R}^3$ is connected [29]. This means one can go from one triangulation to any other by means of repeated number of flips.

2.2 Elements of the theory

2.2.1 The Genotope

As described before, the genotype space G is a set of a^L points in $\mathbb{R}^L$.

Definition 2.2.1. The **Genotope** Π_G is the convex hull of the genotype space.

The convex hull of any finite point configuration is nothing but a convex polytope. In this case, the vertices of the convex polytope represent the a^L possible sequences. For instance, in the 2 loci bi-allelic case, the Genotope is simply a square with vertices (0,0), (0,1), (1,0), (1,1).

Definition 2.2.2. An **allele frequency vector** $\vec{\nu}$ for a bi-allelic population is an L-dimensional vector and its ith entry (ν_i) represents the fraction of the population that has the mutated allele 1 at its ith site.

The vertices of the Genotope can also be viewed as the allele frequency vectors of homogeneous populations composed of only one genotype. Then, each point enclosed by the Genotope represents the allele frequency vector of a heterogeneous population that is composed of several genotypes. Thus, the Genotope is basically a set of all possible allele frequency vectors. Note that the concept of the Genotope can also be extended to diploids.

2.2.2 Triangulations of the Genotope

The Genotope is merely a set of all possibilities. Which possibilities get realised in nature is governed by evolutionary forces. The structure of the fitness landscape encodes one such force, which is the force of natural selection.

Applying the fitness landscape map to the vertices of the Genotope amounts to lifting the configuration to one higher dimension, by raising each vertex by an amount equal to the fitness of the genotype that is represented by that vertex. The convex hull of these raised vertices is also a convex polytope.

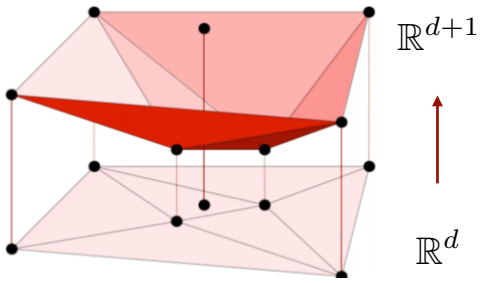


Figure 2.3: An example of a lifting construction that induces regular triangulations, although in this case the lower surface (instead of the upper surface) of the higher dimensional polytope is projected. Adapted from: [28]

The projection of the upper surface of this higher dimensional polytope on the Genotope then gives rise to a triangulation of the Genotope.

This can be formalised as follows: We can extend the definition of the fitness landscape to also assign a fitness value to every allele frequency vector lying inside the Genotope. This can be done by assigning the maximum fitness that a population with the given allele frequency vector can have. This new continuous fitness landscape is a piece-wise linear convex function and the domains of linearity of this function are actually the simplices in the triangulation.

The number of possible triangulations of a given Genotope is finite. For 2 loci, there are 2, for 3 loci there are 74, while for 4 loci, there are already 87959448. Fitness landscapes that induce the same triangulation are said to have the same shape.

Definition 2.2.3. The **shape** of a fitness landscape is the triangulation of the Genotope that is induced by it.

It will become obvious later that landscapes of the same shape have similar epistatic interactions between their loci.

Another important information provided by the shape is that the vertices of the simplex to which the allele frequency vector of a population belongs, are the genotypes that will be present in the **maximally fit population**, given that the allele frequency vector remains fixed during the dynamics. This is not obvious at first glance, but it can be proved using results from linear

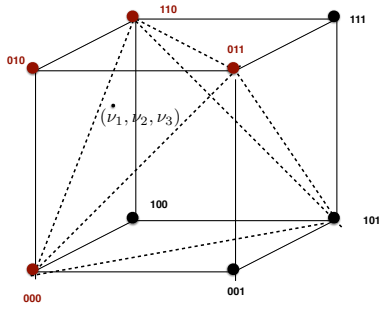


Figure 2.4: A triangulated Genotope

programming[6]. For example, for the allele frequency vector shown in figure 2.4, the maximally fit population will contain the genotypes 000, 011, 110 and 010. This is a useful fact for dynamics like recombination in which the allele frequency vectors remain unchanged. This also implies that the ith entry of the GKZ vector represents the probability that the corresponding genotype occurs in fittest populations conditioned upon allele frequency vectors.

2.2.3 Tools for triangulation

Testing whether a set of subsets of a point configuration comprises a triangulation of that point configuration, is a non-trivial computational problem **[28]**. This is where *circuits* and *Markov bases* come to use. While circuits are combinatorial tools that lead to a fully algorithmic definition of a triangulation, Markov bases exploit the rich link between algebraic geometry and triangulations to construct triangulations. Moreover, these two tools for constructing triangulations reveal patterns of multidimensional epistasis exhibited by the fitness landscape.

In order to better explain these concepts, one needs to introduce the following others:

- Additive epistasis can be measured by looking at linear combinations of genotype fitnesses. Certain sets of these linear combinations form interaction coordinates and both magnitude and sign of these coordinates are relevant when examining the landscape of a biological organism. Now,

[6]Namely, the fundamental theorem of linear programming, which states that the extremal values of a linear function over a convex polytope are attained at its vertices.**[30]**

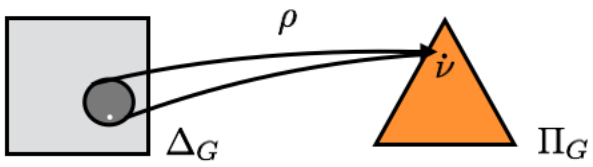


Figure 2.5: A cartoon illustrating the map ρ. Δ_G represents the probability simplex corresponding to the genotype space G.

any fitness landscape can be represented as a vector $\vec{w} \in \mathbb{R}^{|G|}$, where $|G|$ is the number of genotypes. Let $L_G \subset \mathbb{R}^{|G|}$ such that every $\vec{w} \in L_G$ is completely non-epistatic, i.e. there exists an affine-linear form on the Genotope, whose values are w_G at the vertices $\Rightarrow L_G = \{\vec{w} : \vec{v} \cdot \vec{G} + c = w_G \ \forall \vec{G}\}$ where $\vec{v} \cdot + c$ is an affine linear form on Π_G, $\vec{G}$ represents a vertex of Π_G and w_G is the corresponding fitness of vertex $\vec{G}$.

Definition 2.2.4. The **interaction space** is then defined as the dual vector space of the quotient of $\mathbb{R}^G$ modulo L_G i.e. $I_G = (\mathbb{R}^G / L_G)^*$. Thus, elements of I_G are linear forms that vanish on L_G.

Definition 2.2.5. Finally, **circuits** are linear forms that (redundantly) span the interaction space and have non-empty but minimal support. The number of circuits is usually larger than the dimension of I_G ($d(I_G) = 2^L - L - 1$) but is bounded above by $\binom{|G|}{d(I_G)-1}$.

- Let's define ρ to be a map that takes population frequency vectors (that lie in a $2^L - 1$ dimensional simplex) to their corresponding allele frequency vectors (that are contained in the Genotope) i.e. $\rho : \vec{x} \mapsto \vec{\nu}$, where $\vec{x} = (x_{00\ldots0}, \ldots., x_{11\ldots1})$ and $\vec{\nu} = (\nu_1, \ldots., \nu_L)$. This map is clearly not a bijection and therefore the pre-image of any $\vec{\nu} \in$ Genotope (Π_G) is $\rho^{-1}(\vec{\nu}) = \{\vec{x} : \rho(\vec{x}) = \vec{\nu}\}$. This is illustrated in figure 2.5. The dimensions of $\rho^{-1}(\vec{\nu}) = 2^L - L - 1 = d(I_G)$. ρ when written as a matrix, turns out to be the matrix whose columns are the vertices of the Genotope.
  For instance, in the 2 loci case:
  $\rho = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix}$,
  while in the 3 locus case:
  $\rho = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}$.

- The **integral kernel** of ρ i.e. $\ker_{\mathbb{Z}}(\rho) = \{\vec{k} : \rho(\vec{k}) = 0\} \cap \mathbb{Z}^{2^L-1}$ defines the interaction space for the genotypes. The Markov basis or the circuits are a non-independent basis for the interaction space. I will henceforth omit the subscript of $\ker_{\mathbb{Z}}(\rho)$, however unless otherwise stated, $\ker(\rho)$ will still refer to the integral kernel and not the entire kernel.

- A more concrete definition of Markov bases exists in the context of Toric ideals.

  Definition 2.2.6. A Toric ideal, $I_\rho = \langle p^{\vec{u}} - p^{\vec{v}} : \rho(\vec{u}) = \rho(\vec{v})\rangle$, where, $p^{\vec{a}} = p_1^{a_1} p_2^{a_2} ... p_n^{a_n}$ represents a monomial in n variables $p_1, p_2 ... p_n$ and $\langle P\rangle$ represents the ideal[7] generated by a set of polynomials P. In other words, a Toric ideal is the ideal generated by binomials of the form $p^{\vec{u}} - p^{\vec{v}}$ where $\vec{u}$ and $\vec{v}$ satisfy the above mentioned property.

  Definition 2.2.7. A finite set of binomials, with the above stated property, that generates the Toric ideal I_ρ is called a **Markov basis** for the Toric ideal, i.e. if $I_\rho = \langle\{x^{\vec{m}_+} - x^{\vec{m}_-} : \rho(\vec{m}_+) = \rho(\vec{m}_-)\}\rangle$ then, the finite set of all $\vec{m} = \vec{m}_+ - \vec{m}_-$ is called a Markov basis.

- From this definition, it can be seen that $\vec{m} \in \ker(\rho) \because \rho(\vec{m}_+) = \rho(\vec{m}_-) \Rightarrow \rho(\vec{m}_+ - \vec{m}_-) = \rho(\vec{m}) = 0$. Therefore, a Markov basis can alternatively be defined as a subset $\boldsymbol{B}$ of $\ker(\rho)$ that satisfies the following conditions:

  1. If $\forall \vec{u}, \vec{v}$ satisfying $\rho\vec{u} = \rho\vec{v}$, $\exists\{\vec{m}_i\}_{i=1}^{l}$, such that $\vec{u} + \sum_i \vec{m}_i = \vec{v}$ and
  2. $\forall j$ satisfying $1 \le j \le l$, $\vec{u} + \sum_{i=1}^{j} \vec{m}_i \ge 0$

  If conditions 1. and 2. are met, then $\boldsymbol{B}$ is a Markov basis and every $m \in \boldsymbol{B}$ is called a **move**. Markov basis can be used to do Monte-Carlo simulations. For example, in our context, adding a move to a probability vector will give a new probability vector with the same allele frequency vector as the original probability vector. This way, one can hop in the subset of the probability simplex which maps to a particular allele frequency vector. Further, with some pre-knowledge about the stationary probability distribution, one can get the probability of taking a certain step. This can then be used to compute equilibrium averages of quantities.

- Lastly, circuits measure additive epistasis while Markov basis elements measure multiplicative epistasis.

[7] For more information on polynomial rings and ideals, see [27]

2.3 Examples of shapes

2.3.1 2 locus case

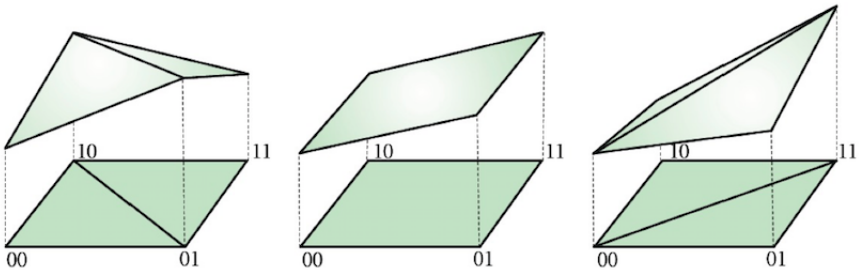


Figure 2.6: Possible shapes for 2 loci landscapes. Landscapes with $e < 0$ have the shape shown in the left most figure, while those with $e > 0$ have the shape shown in the right most figure. The central figure corresponds to non-epistatic landscapes that do not triangulate the Genotope and hence have no shape.

Triangulations for the 2-loci Genotope are almost trivial because there is only one possible interaction between the 2 loci. This interaction is measured by the circuit: $e = w_{00} + w_{11} - w_{01} - w_{10}$. This circuit gives rise to two shapes that are shown in figure 2.6.

The shape of a two locus fitness landscape is not very informative about its topography because the probability of exhibiting a particular type of sign epistasis (either simple, reciprocal or no sign epistasis) remains independent of the shape of the landscape. This is because for example:

$$P(\text{reci}) = P(\text{reci} \mid \text{shape1})P(\text{shape1}) + P(\text{reci} \mid \text{shape2})P(\text{shape2}) \tag{2.2}$$

where, $P(\text{reci})$ represents the probability of having reciprocal sign epsiatsis. Now, for HoC fitness landscapes $P(\text{shape1}) = P(\text{shape2}) = 0.5$.

$$\Rightarrow P(\text{reci}) = 0.5 \cdot P(\text{reci} \mid \text{shape1}) + 0.5 \cdot P(\text{reci} \mid \text{shape2}) \tag{2.3}$$

Now, $P(\text{reci} \mid \text{shape1}) = P(\text{reci} \mid \text{shape2})$ because the two shapes differ merely by labelling. Exchanging the labels $w_{11} \longleftrightarrow w_{10}$ and $w_{00} \longleftrightarrow w_{01}$ exchanges the shape as well.

$$\Rightarrow P(\text{reci} \mid \text{shape}) = P(\text{reci}) \tag{2.4}$$

All we can know from the shape is the location of the two peaks, given that there is reciprocal sign epistasis. This however is not surprising because

in the 2 loci case, shapes are distinguished only by the sign of a circuit measuring magnitude epistasis. Therefore, the more useful knowledge about sign epistasis (and thus the number of peaks) is not contained in the shapes of the 2-loci landscape.

2.3.2 3 locus case

As previously mentioned, there are 74 possible triangulations for the bi-allelic, 3 locus case. These belong to **6** symmetry classes or interaction **types**. I will explain why there are 74 shapes and 6 types, through a small story about *Newton polytopes* and *hyperdeterminants*. This story has been completely borrowed from **[31]**.

- The **Newton polytope** N(G) of a polynomial G is the convex hull of the exponent vectors of the monomials which appear in the expansion of G.
- The **hyperdeterminant** of a $2\times2\times2$-tensor, D_{222} is an irreducible polynomial in eight variables with twelve monomials of degree four and is called a *tangle* in physics literature. Note that it is the higher dimensional analog of the determinant of a 2×2 matrix which is the polynomial $D_{22} = x_{00}x_{11} - x_{01}x_{10}$ in four variables.
- **N(D_{222})** is the convex hull in $\mathbb{R}^8$ of the six rows of the following matrix:

$$\begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 \\ 2 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \\ 0 & 2 & 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 2 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 & 0 & 0 & 0 \end{bmatrix}$$

  This is because the exponents of these monomials are vertices of N(D_{222}). This makes them *extreme monomials*. The exponents of the remaining 6 monomials lie in the interior of N(D_{222}) and do not contribute to the convex hull. The *f-vector* records the number of faces of dimension 0,1,2...d-1 and f(N(D_{222}))=(6,14,16,8), meaning that N(D_{222}) has 6 vertices, 14 edges, 16 2-dimensional faces and 8 3-dimensional faces.

- A final relevant character in the story is the *principal determinant* of the 3-cube i.e.

$$\begin{aligned} E_{222} = D_{222} \cdot (x_{000}x_{011} - x_{001}x_{010}) \cdot (x_{000}x_{101} - x_{001}x_{100}) \cdot \\ (x_{000}x_{110} - x_{010}x_{100}) \cdot (x_{001}x_{111} - x_{011}x_{101}) \cdot \\ (x_{010}x_{111} - x_{011}x_{110}) \cdot (x_{100}x_{111} - x_{101}x_{110}) \cdot \\ x_{000} \cdot x_{001} \cdot x_{010} \cdot x_{011} \cdot x_{100} \cdot x_{101} \cdot x_{110} \cdot x_{111}. \end{aligned}$$

  This is a polynomial of degree 24 with 231 monomials out of which 74 are extreme monomials. It turns out that N(E_{222}) is the secondary polytope of the 3 cube. It is 4-dimensional and its *f-vector* is (74,152,100,22). Further, its 74 vertices are in bijection with the regular triangulations of 3 cube. Moreover, the symmetry group of the 3-cube is the Weyl group B_3 of order 48 and it turns out that the 74 extreme monomials come in 6 *orbits*[8]. And that solves the mystery of 74 shapes and 6 types!

The *tight spans* of the 6 types of shapes are shown in figure 2.7.
The following is a brief description of the types:

1. **Type 1** contains 2 shapes that divide the cube into five tetrahedra, one central tetrahedron of normalized volume two surrounded by four of normalized volume one.

2. **Type 2** contains 8 shapes that are generated by slicing off the three vertices adjacent to a fixed vertex and cutting the remaining bipyramid into three tetrahedra.

3. **Type 3** contains 24 shapes that are generated by picking a diagonal and two of the other six vertices that are diagonal on a facet, and slicing them off.

4. **Type 4** has 12 shapes which are indexed by ordered pairs of diagonals. The end points of the first diagonal are sliced off, and the remaining octahedron is triangulated using the second diagonal.

5. **Type 5** has 24 shapes that are indexed by a diagonal and one other vertex which is sliced off, and the remaining polytope is divided into a pentagonal ring of tetrahedra around the diagonal.

6. **Type 6** has 4 shapes that are indexed by the diagonals. The cube is divided into a hexagonal ring of tetrahedra around the diagonal.

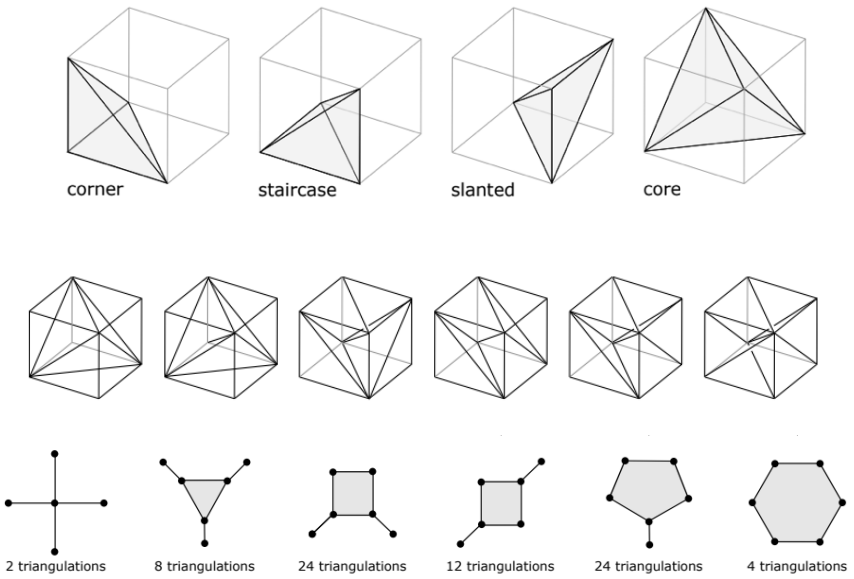


Figure 2.7: Top to bottom: The types of tetrahedra that appear in the triangulation of the 3 cube, the six types of triangulations of the 3 cube and the tight spans of the 6 types– the vertices represent the tetrahedra in the triangulation and two vertices are connected if the tetrahedra share a common triangle. Source: **[32]**

The distributions of shapes and types of HoC fitness landscapes are shown in figures 2.9 and 2.8 for the uniform distribution and the exponential distribution. As is evident, the distribution depends upon the probability distribution from which the fitness values are assigned.

For uniformly distributed HoC landscapes, each shape of a particular type is equally likely to occur. This is because shapes of a particular type can be obtained from one another by re-labelling of the indices. Moreover, one could naively expect each type to be equally likely to occur. This however cannot be the case because the probabilities would not be normalised any more. Type 3 and type 5 are actually the most abundant types with an abundance of approximately 25% each. The other four types have an abundance of nearly 12.5% each. Another way to obtain the shape distribution and verify the re-

[8]For a group G that acts on a set X, the orbit of every x ∈ X is $\mathrm{Orb}_x = \{g.x : g \in G\} \subset X$.

sults obtained by triangulating the fitness landscape is to look at the circuits. The 20 circuits for the 3-loci case are as follows:

$$
\begin{aligned}
a &:= w_{000} - w_{010} - w_{100} + w_{110} \\
b &:= w_{001} - w_{011} - w_{101} + w_{111} \\
c &:= w_{000} - w_{001} - w_{100} + w_{101} \\
d &:= w_{010} - w_{011} - w_{110} + w_{111} \\
e &:= w_{000} - w_{001} - w_{010} + w_{011} \\
f &:= w_{100} - w_{101} - w_{110} + w_{111} \\
g &:= w_{000} - w_{011} - w_{100} + w_{111} \\
h &:= w_{001} - w_{010} - w_{101} + w_{110} \\
i &:= w_{000} - w_{010} - w_{101} + w_{111} \\
j &:= w_{001} - w_{011} - w_{100} + w_{110} \\
k &:= w_{000} - w_{001} - w_{110} + w_{111} \\
l &:= w_{010} - w_{011} - w_{100} + w_{101} \\
m &:= w_{001} + w_{010} + w_{100} - w_{111} - 2w_{000} \\
n &:= w_{011} + w_{101} + w_{110} - w_{000} - 2w_{111} \\
o &:= w_{010} + w_{100} + w_{111} - w_{001} - 2w_{110} \\
p &:= w_{000} + w_{011} + w_{101} - w_{110} - 2w_{001} \\
q &:= w_{001} + w_{100} + w_{111} - w_{010} - 2w_{101} \\
r &:= w_{000} + w_{011} + w_{110} - w_{101} - 2w_{010} \\
s &:= w_{000} + w_{101} + w_{110} - w_{011} - 2w_{100} \\
t &:= w_{001} + w_{010} + w_{111} - w_{100} - 2w_{011}
\end{aligned}
$$

Circuits a-f check for pairwise epistasis, g-l for marginal epistasis between two pairs of loci and m-t for three way epistasis in relation to total pairwise epistasis.

The shapes are characterized by the sign pattern of a selected number of circuits. For instance, for a fitness landscape to have shape 1, circuits t, q, o and m must be positive. Thus, the shape abundances can also be computed by looking at probabilities that a random landscape will have a circuit sign pattern that characterises that particular shape. One more motivation for computing the abundances differently was to check if the abundances are rational numbers (i.e. 1/4 and 1/8).

Since circuits are nothing but linear combinations of i.i.d random variables, one can estimate the probability of obtaining a set of 8 random vari-

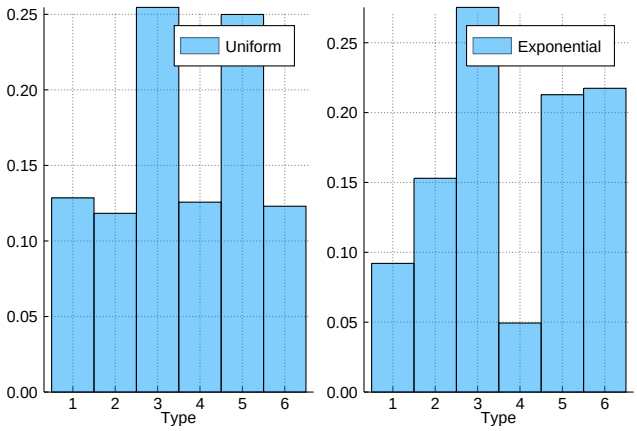


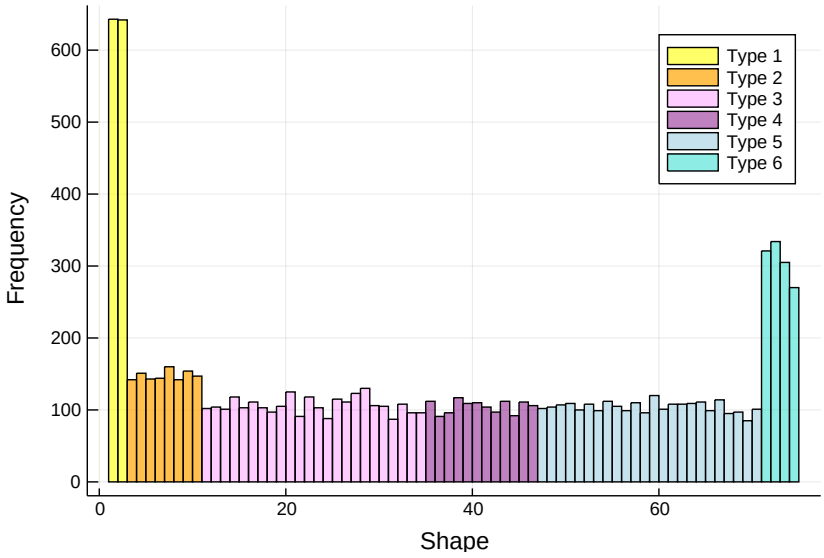
Figure 2.8: The relative abundances of the 6 types for HoC landscapes generated from a uniform distribution and an exponential distribution.

ables that satisfy a certain circuit sign pattern by computing the volume of the polytope bounded by hyper-planes given by the circuit sign pattern, e.g.

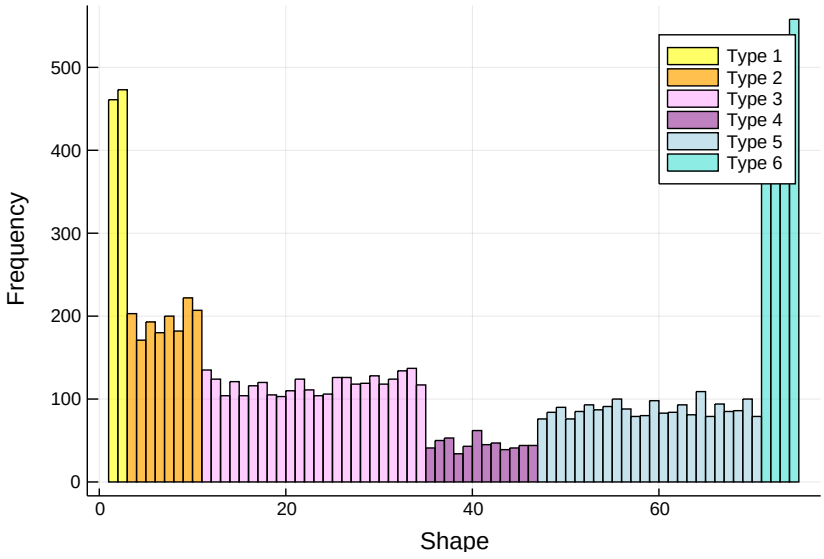
$$P(\text{shape1}) = \int_0^1 \ldots \int_0^1 \Pi_{i=1}^{8} dw_i \Theta(t(\{w_i\}))\Theta(q(\{w_i\}))\Theta(o(\{w_i\}))\Theta(m(\{w_i\})) \quad (2.5)$$

These integrals were easily computed by using the Monte-Carlo method. The resultant abundances matched the previously obtained ones (shown in table 3.1), however it still could not be ascertained whether these abundances were rational numbers.

Finally, figure 2.10 summarises the GKZ vectors, defining circuits and neighbours on the secondary polytope of all the 74 shapes.



(a) Shape distribution for HoC landscapes with uniformly distributed fitness values.



(b) Shape distribution for HoC landscapes with exponentially distributed fitness values.

Figure 2.9

#/T	GKZ	Out-edges	#/T	GKZ	Out-edges
1/1	15515115	3t4q5o6m	38/4	31355313	39$\bar{l}$44$\bar{g}$51c59d
2/1	51151551	7s8r9p10n	39/4	31533513	38l44$\bar{i}$53e60f
3/2	14436114	1$\bar{t}$11b13d17$\bar{e}$	40/4	33155133	42$\bar{j}$45$\bar{g}$54a61b
4/2	14614314	1$\bar{q}$12b14f18$\bar{c}$	41/4	33511533	43$\bar{h}$46$\bar{i}$55a62b
5/2	16414134	1$\bar{o}$15d16f19$\bar{a}$	42/4	35133153	40j45$\bar{k}$57e63f
6/2	34414116	1$\bar{m}$28$\bar{e}$29$\bar{c}$31$\bar{a}$	43/4	35311353	41h46$\bar{k}$58c64d
7/2	41163441	2$\bar{s}$20a22c26$\bar{f}$	44/4	51333315	38g39i65$\bar{b}$68$\bar{a}$
8/2	41341641	2$\bar{r}$21a23e27$\bar{d}$	45/4	53133135	40g42k66$\bar{d}$69$\bar{c}$
9/2	43141461	2$\bar{p}$24c25e30$\bar{b}$	46/4	53311335	41i43k67$\bar{f}$70$\bar{e}$
10/2	61141443	2$\bar{n}$32$\bar{f}$33$\bar{d}$34$\bar{b}$	47/5	13356222	11$\bar{d}$13$\bar{b}$35f71$\bar{e}$
11/3	13446213	3$\bar{b}$12$\bar{l}$47d51$\bar{e}$	48/5	13623522	12$\bar{f}$14$\bar{b}$36d72$\bar{c}$
12/3	13624413	4$\bar{b}$11l48f53$\bar{c}$	49/5	16323252	15$\bar{f}$16$\bar{d}$37b73$\bar{a}$
13/3	14346123	3$\bar{d}$15$\bar{j}$47b54$\bar{e}$	50/5	22265331	20$\bar{c}$22$\bar{a}$35e71$\bar{f}$
14/3	14613423	4$\bar{f}$16$\bar{h}$48b55$\bar{c}$	51/5	22356213	11e17$\bar{b}$38$\bar{c}$71d
15/3	16324143	5$\bar{d}$13j49f57$\bar{a}$	52/5	22532631	21$\bar{e}$23$\bar{a}$36c72$\bar{d}$
16/3	16413243	5$\bar{f}$14h49d58$\bar{a}$	53/5	22623513	12c18$\bar{b}$39$\bar{e}$72f
17/3	23346114	3e28$\bar{g}$51b54d	54/5	23256123	13e17$\bar{d}$40$\bar{a}$71b
18/3	23613414	4c29$\bar{i}$53b55f	55/5	23612523	14c18$\bar{f}$41$\bar{a}$72b
19/3	26313144	5a31$\bar{k}$57d58f	56/5	25232361	24$\bar{e}$25$\bar{c}$37a73$\bar{b}$
20/3	31264431	7$\bar{a}$21$\bar{l}$50c59$\bar{f}$	57/5	26223153	15a19$\bar{d}$43$\bar{e}$73f
21/3	31442631	8$\bar{a}$20l52e60$\bar{d}$	58/5	26312253	16a19$\bar{f}$43$\bar{c}$73d
22/3	32164341	7$\bar{c}$24$\bar{j}$50a61$\bar{f}$	59/5	31265322	20f26$\bar{a}$38$\bar{d}$71c
23/3	32431641	8$\bar{e}$25$\bar{h}$52a62$\bar{d}$	60/5	31532622	21d27$\bar{a}$39$\bar{f}$72e
24/3	34142361	9$\bar{c}$22j56e63$\bar{b}$	61/5	32165232	22f26$\bar{c}$40$\bar{b}$71a
25/3	34231461	9$\bar{e}$23h56c64$\bar{b}$	62/5	32521632	23d27$\bar{e}$41$\bar{b}$72a
26/3	41164332	7f32$\bar{g}$59a61c	63/5	35132262	24b30$\bar{c}$42$\bar{f}$73e
27/3	41431632	8d33$\bar{i}$60a62e	64/5	35221362	25b30$\bar{e}$32$\bar{d}$73c
28/3	43324116	6e17g65$\bar{c}$66$\bar{a}$	65/5	52323216	28c29e44b74$\bar{a}$
29/3	43413216	6c18i65$\bar{e}$67$\bar{a}$	66/5	53223126	28a31e45d74$\bar{c}$
30/3	44131362	9b34$\bar{k}$63c64e	67/5	53312226	29a31c46f74$\bar{e}$
31/3	44313126	6a19k66$\bar{e}$67$\bar{c}$	68/5	61232325	32d33f44a74$\bar{b}$
32/3	61142334	10f26g68$\bar{d}$69$\bar{b}$	69/5	62132235	32b34f45c74$\bar{d}$
33/3	61231434	10d27i68$\bar{f}$70$\bar{b}$	70/5	62221335	33b34d46e74$\bar{f}$
34/3	62131344	10b30k69$\bar{f}$70$\bar{d}$	71/6	22266222	47e50f51$\bar{d}$54$\bar{b}$59$\bar{c}$61$\bar{a}$
35/4	13355331	36$\bar{l}$37$\bar{j}$47$\bar{f}$50$\bar{e}$	72/6	22622622	48c52d53$\bar{f}$55$\bar{b}$60$\bar{e}$62$\bar{a}$
36/4	13533531	35l37$\bar{h}$48$\bar{d}$52$\bar{c}$	73/6	26222262	49a56b57$\bar{f}$58$\bar{d}$63$\bar{e}$64$\bar{c}$
37/4	15333351	35j36h49$\bar{b}$56$\bar{a}$	74/6	62222226	65a66c67e68b69d70f

Figure 2.10: All 74 shapes of the 3-cube with their GKZ vectors and the circuit sign patterns that they show; a means circuit $a > 0$ while $\bar{a}$ means circuit $a < 0$. Also, mentioned are the neighbouring shapes that differ only by the sign of one particular circuit.

Chapter 3

Shapes and their contemporaries

Whenever a new theory is developed, it becomes important to assess its usefulness and to also compare it to pre-existing theories. This is what I strive to do in this chapter.

3.1 Applications of shapes

So far, it seems that this new theory can be interesting in the following areas:

1. **Analysing empirical data**: As was mentioned previously, shape theory helps in studying all possible interactions between a given set of loci. This enables a more fine scaled study of the interactions in empirical fitness landscapes. Further, triangulating empirical landscapes can give information about the composition of fittest populations– a fact that can be tested experimentally. It can also reveal which genotypes are "sliced off" in the triangulation. Moreover, shape theory is also applicable to combinatorially incomplete, multi-locus landscapes. This is useful because for long sequences ($L > 20$), not all genotypes are realized in nature.

2. **Studying purely recombining populations**: Allele frequency preserving dynamics, like recombination, can be studied only on a subset of the population simplex, which is given by $\rho^{-1}(\vec{v})$.

3. **Studying evolution with mutation and selection**: One can study what the shape tells us about general evolutionary processes like deterministic mutation-selection dynamics.

4. **Studying the evolution of recombination**: It is known that the deterministic evolution of recombination depends upon the epistatic interactions between the loci [33]. Since shapes are a way of classifying the various types of possible interactions, it can be interesting to study whether a particular shape opposes or supports the evolution of recombination.

The last two applications are only valid for 2 and 3 locus landscapes because there are too many shapes for 4 and more loci. For the sake of completeness, I'll mention that this theory has also been used to compute the human Genotope, in order to describe the shapes of landscapes associated with measurements of phenotypes across populations [34].

In the remaining part of this chapter, shapes of three locus landscapes are compared to graphs, the Walsh spectrum and the γ measure that was introduced in [25].

3.2 Shapes in comparison to graphs

I compared shapes with graphs in three different contexts:

Type	Abundance	Reciprocal SE	Simple SE	No SE
1	0.12	3.68	1.66	0.66
2	0.12	2.57	1.85	1.58
3	0.24	1.86	2.04	2.1
4	0.13	1.49	2.25	2.26
5	0.25	1.6	1.96	2.43
6	0.12	1.46	2.27	2.28

Table 3.1: Column two shows the relative abundance of each of the 6 types, for HoC landscapes with uniformly distributed fitness values. The remaining columns show the average number of reciprocal, simple and no sign epistasis (SE) motifs in representative landscapes of the 6 types.

1. **Ruggedness of fitness landscapes:** In order to compare shapes with graphs, for HoC landscapes of each type (with uniformly distributed fitness values), I counted the number of times reciprocal sign epistasis, simple sign epistasis and no sign epistasis motifs occur. The sum of the number of these occurrences must add up to 6, because there are 6 faces

of the cube. Table 3.1 summarises these results. Here, unlike the 2-loci case, the types favour certain motifs more than others. In other words, $P(\text{reci} \mid \text{type})$ is no longer independent of the type (and thus the shape). Type 1 landscapes are very likely to show reciprocal sign epistasis and should thus be quite rugged **[14]**. The probability to exhibit reciprocal sign epistasis nearly monotonically decreases with the type (type 4 being an exception).

The difference in occurrence of various sign epistasis motifs immediately tells us that the number of peaks will also differ between the landscapes of various types. Results relating to the number of peaks are shown in figure 3.1. The mean number of peaks shows a trend similar to the probability of occurrence of reciprocal sign epistasis. While type 1 landscapes have nearly three peaks on average, type 6 have a little less than 2.

While graphs unequivocally tell us about the number of peaks in the landscape, shapes give rise to distributions of number of peaks. Given that the evolutionary dynamics (e.g. length of adaptive walks) and the stationary state (if it exists) depend strongly on the number of peaks, shapes cannot be of as much use in tackling problems related to adaptive walks.

2. **Experimental applications:** In the context of experiments, it is easier to determine the fitness orders than the exact fitness values. Moreover, it was recently shown in **[24]** that information about higher order epistasis can also be obtained from graphs. This is good news because often partial orders are the best one can expect from experimental measurements. That said, shapes too can be used to study partially determined fitness landscapes but merely the ordering of fitnesses is not sufficient– knowing the absolute values of those fitnesses is a prerequisite. Also, often times, only the knowledge of whether or not a landscape has higher order epistasis is not enough, one must also know the strength of that epistasis in order to infer something about population dynamics (as will be seen in chapter 5). This information about the strength of epistasis is contained in the circuits or the Markov bases of the interaction space but not in graphs. Moreover, fitness orders and consequently graphs are not enough to compute the shapes of three locus landscapes **[24]**. This is not surprising given that absolute fitness values are required to compute shapes while only partial fitness orders to compute graphs.

3. **Classifying fitness landscapes:** Finally, when it comes to segregating

multi locus landscapes, the number of both shapes and graphs grows hopelessly. As opposed to 74 shapes for the three locus case that fall into 6 types, there are 1,862 fitness graphs of 54 types.

It is important to note that graphs and shapes are not opposing viewpoints but complementary. Kristina Crona nicely sums up this comparative study of graphs and shapes by saying, "...graphs provide information that cannot be obtained from the geometric classification, and vice versa..." **[35]**.

3.3 Shapes in comparison to the Walsh spectra

As previously mentioned, the Walsh coeffiecients can be calculated from a fitness landscape by a linear transformation i.e. $\vec{e} = \hat{V} \cdot \hat{H}_L \cdot \vec{w}$, where $\vec{w}$ is the vector of the fitness values ordered by the binary number that the corresponding bi-allelic sequence represents, $\vec{e}$ is the vector of the Walsh coefficients, $\hat{H}_L$ is the Hadamard matrix of order 2^L and $\hat{V}$ is a diagonal matrix for the purpose of normalisation. The epistatic order of the Walsh coeffcients is also determined by the binary number to which that coefficient corresponds e.g. e_3 is the third element of $\vec{e}$ so it corresponds to the binary number 011 and represents the second order (pairwise) interaction between loci 2 and 3.

The Walsh coefficients are actually intimately connected to circuits. The coefficents of order≥ 2 form a basis of the interaction space (and are referred to as interaction coordinates in **[26]**). However, they can be expressed as linear combinations of circuits, which as previously mentioned, span the interaction space, e.g. $e_8 = b - a$. Thus, circuits contain more fine scaled information than Walsch coefficients. Moreover, for combinatorially incomplete landscapes, circuits are more canonical than interaction coordinates.

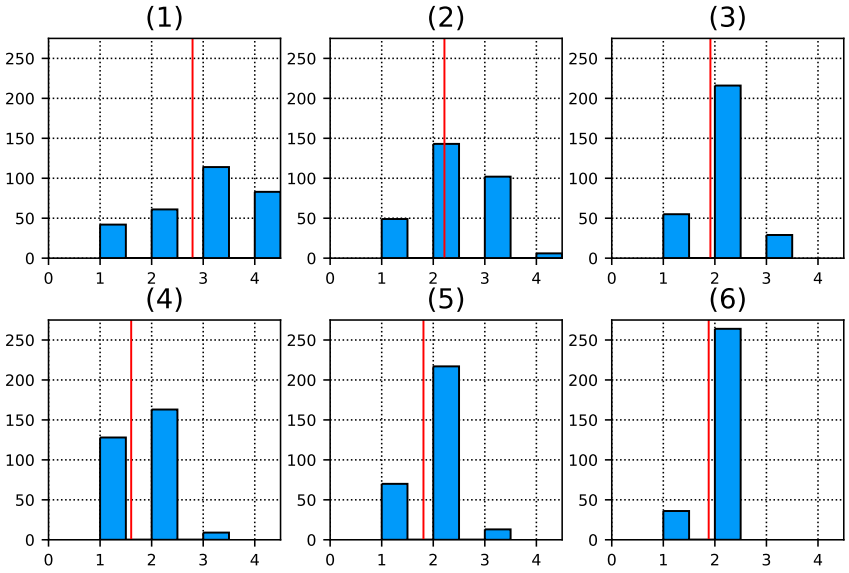
Now, the contribution of the nth epistatic order can be summarised by $W_n = \sum e_j^2$ where e_j represents the elements of $\vec{e}$ corresponding to nth order epistasis. Then,

$$F_{\text{total}} = \frac{\sum_{n=2}^{L} W_n}{\sum_{n=1}^{L} W_n} \tag{3.1}$$

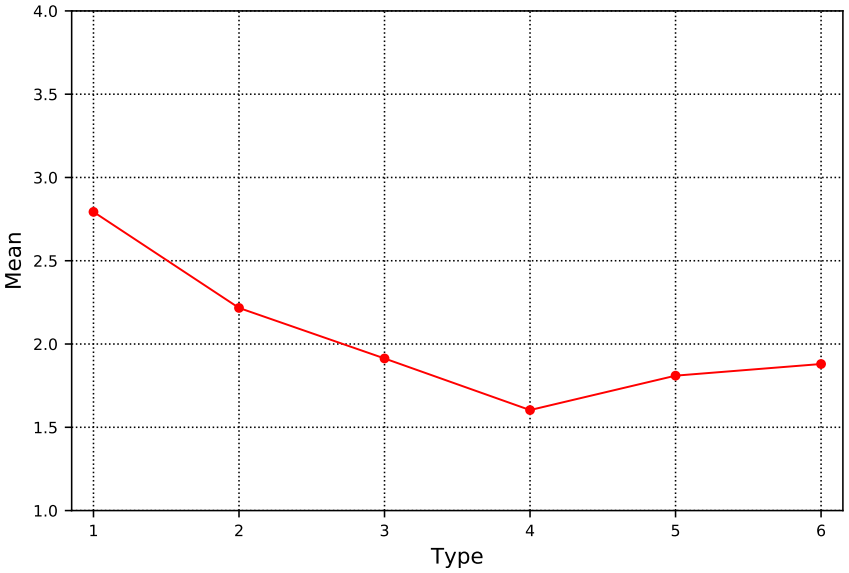
represents the total fraction of epistatic contribution. Thus, $F_{\text{total}} = 0$ for additive landscapes while $F_{\text{total}} \longrightarrow 1$ as $L \longrightarrow \infty$ for HoC landscapes. Similarly,

$$F_{\text{high}} = \frac{\sum_{n=3}^{L} W_n}{\sum_{n=1}^{L} W_n} \tag{3.2}$$

represents the contribution of solely higher order epistasis and excludes pairwise epistasis.

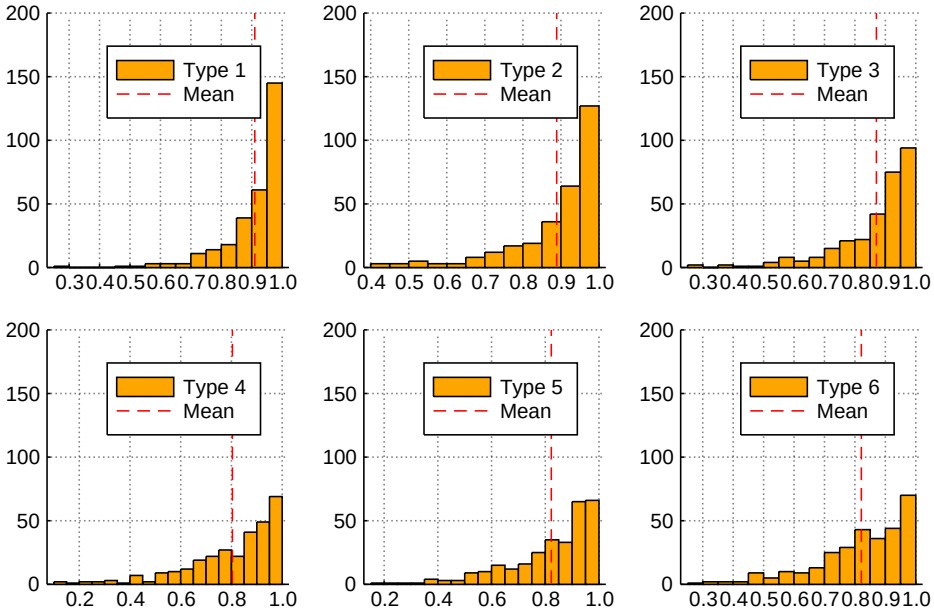


(a) (1-6): Distributions of number of peaks for each of the six types of landscapes

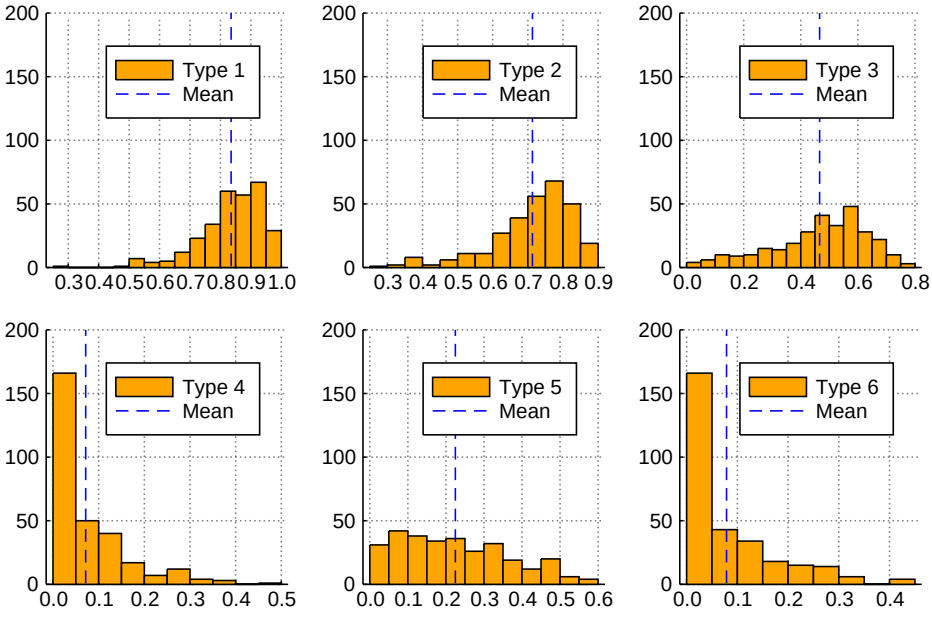


(b) Mean number of peaks as a function of the type

Figure 3.1: Topography of landscapes of different types.



(a) The distribution of F_{total} for landscapes of each of the six types.



(b) The distribution of F_{high} for landscapes of each of the six types.

Figure 3.2: Comparison of shapes with Walsh coefficients.

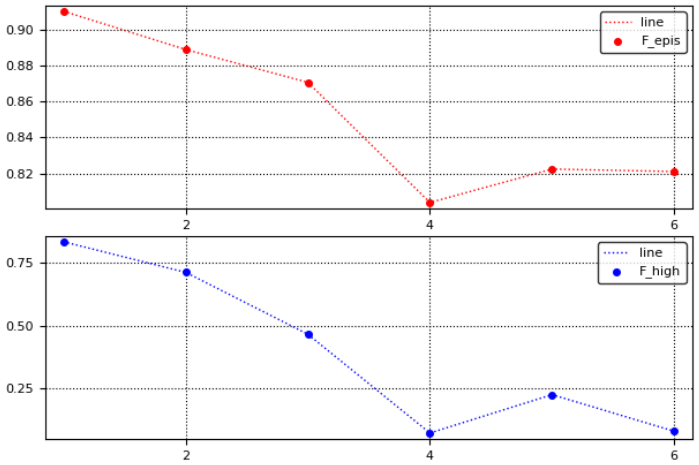


Figure 3.3: Comparison of the means of F_{total} and F_{high} for all the 6 types.

I investigated these two quantities for landscapes of each of the 6 types. The results are shown in figures 3.2 and 3.3. Since these are HoC landscapes, as expected, F_{total} is very close to one for all the types, however the mean becomes smaller as we go from type 1 to type 6, indicating that type 6 landscapes have relatively greater additive contribution than type 1 landscapes. A more interesting trend is observed for F_{high} distributions. For types 1-3, the distribution is peaked close to 1, while for types 4-6 the peak shifts to a value close to zero. This indicates that types 1-3 show greater higher order epistasis than types 4-6. In **[26]**, it was indicated that type 1 is likely to either show very high or very low higher order epistasis, however on average its F_{high} is still larger than all the other types. In fact, the mean contribution to higher order epistasis reduces as we go from type 1 to type 6. The decline is strikingly similar to what is observed for F_{total}, but the magnitude of the decline is much greater for F_{high}. Not too surprisingly, this trend is also correlated with the trend seen for the mean number of peaks in the previous section (figure 3.1). Further, the minimum number of peaks, total sign epistasis and higher order epistasis, on average, consistently occur at type 4.

To summarize, the comparison with the Walsh spectra furthers our intuition about what kind of landscapes are encompassed by each of the 6 types.

3.4 Shapes in comparison to the γ measure

The γ measure measures the correlation of mutational effects on different backgrounds. As mentioned before, $\gamma = \text{Cor}(s(g), s(g_1))$. Since it is a correlation,

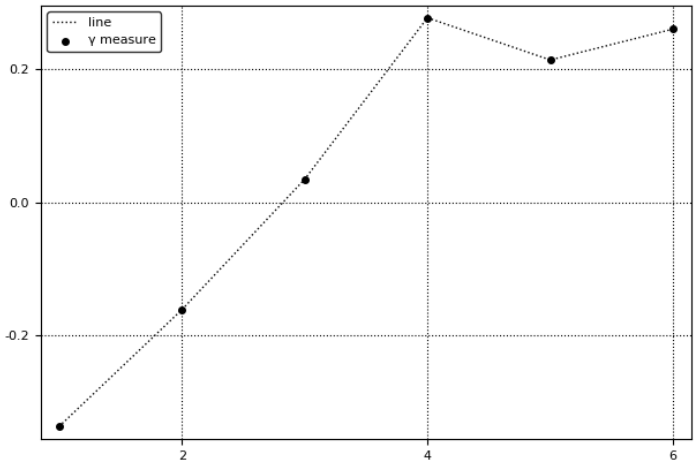


Figure 3.4: The mean of the γ measure for landscapes of each of the six types.

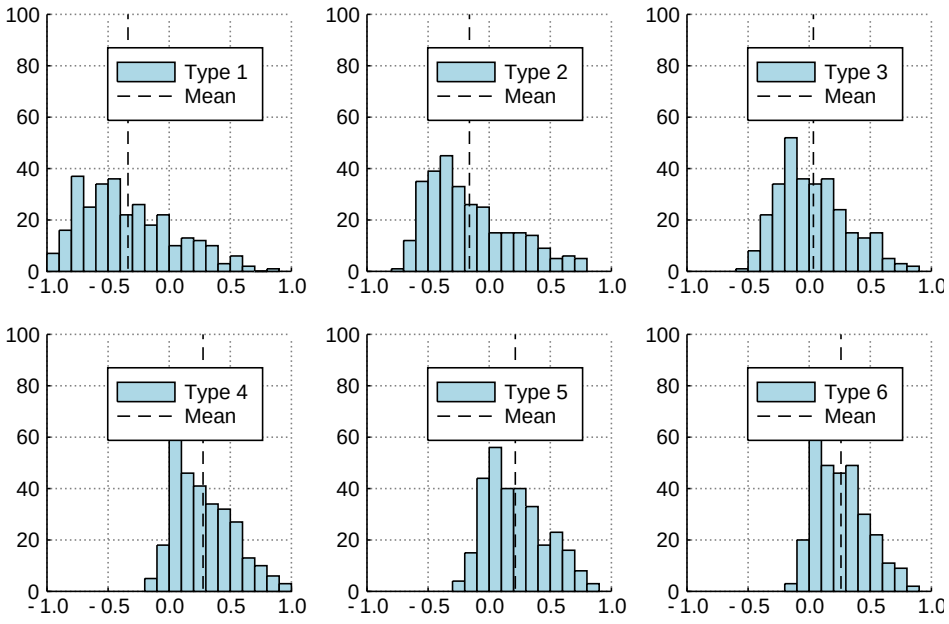


Figure 3.5: The distribution of the γ measure for landscapes of each of the six types.

$-1 \leq \gamma \leq 1$. Consequently, landscapes with magnitude epistasis have $0 \leq \gamma < 1$, landscapes with simple sign epistasis have $-1/3 \leq \gamma < 1$ and landscapes with reciprocal sign epistasis have $-1 \leq \gamma < 0$.

The results for landscapes belonging to different types are shown in figures

3.4 and 3.5.

For landscapes generated by the NK model, $E[\gamma] \simeq 1 - \frac{K}{L-1}$ **[25]**. Now for HoC landscapes, $K = L - 1 \Rightarrow E[\gamma] \simeq 0$. For additive landscapes, $\gamma = 1$. Finally, for landscapes with maximal number of peaks[1], $\gamma = -1$. So in some sense, γ and F_{total} measure opposite effects. This explains why the plots of their means versus types (figures 3.3 (top) and 3.4) look like mirror images. Moreover, the results shown in figure 3.5 basically reinforce the fact that landscapes of type 6 are on average more correlated than those of type 1. The mean of γ is negative for types 1 and 2. This indicates the presence of sign and reciprocal sign epistasis motifs and agrees with what was seen in the section on fitness graphs. Moreover, for types 4-6, γ does not exceed 0.3. This indicates the departure from additivity due to magnitude epistasis.

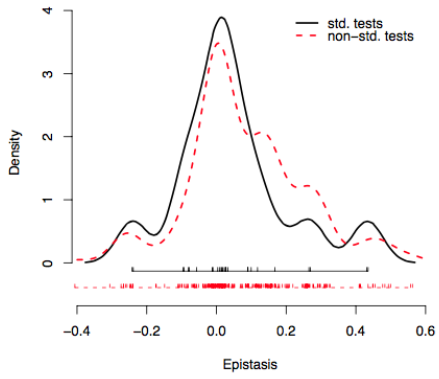
[1] These are called egg-box landscapes

Chapter 4

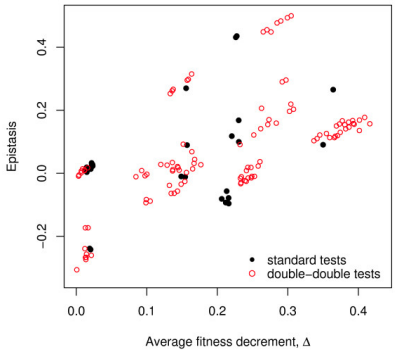
Application to empirical landscapes

4.1 Previous work

In an accompanying paper [36], Beerenwinkel et al exhibited the utility of their shape theory in analysing empirical data. They analysed pre-existing data of Elena and Lenski and showed that the new theory gave more insights into gene interactions in the fitness landscape than were previously known.



(a) The distribution of standard and non-standard tests



(b) The correlation of epistasis and the effect size of the mutations

Figure 4.1: Main results of [36]

They looked at a partial 9 locus landscape with 9 single mutants (all deleterious) and a restricted set of only 27 double mutants. Triple or higher mu-

tants were not considered in their analysis. As mentioned before, the Markov basis for a landscape is a non-independent basis of the interaction space. Put differently, it is a generalisation of the concept of pairwise epistasis. For their landscape, the Markov basis comprised of 243 elements (or tests of interactions) including 27 standard tests (i.e. pairwise tests) and 216 non-standard tests. Although the non-standard tests were dependent (and this must be accounted for in the conclusions), the authors claimed that it was important to consider these tests as well. The reason for that was that the non-standard tests spanned greater Hamming distances than 2. In some sense, epistasis is implicitly assumed to be predominantly due to pairwise interactions and thus, measuring epistasis between pairs of loci (i.e. between genotypes that constitute vertices of the facets of the hypercube) appears intuitive. This is perhaps why Elena and Lenski too, had only looked at the smaller subset of 27 pairwise tests. However, there is no prima facie reason to neglect the non-standard tests of higher order epistasis. Further, some of these non-standard tests showed that certain mutations "mix" better than others, in the sense that they do not cancel the effect of the other mutation with which they occur. This difference in "mixability" causes some elements of the Markov basis to be more likely to have a particular sign and as a result, such landscapes are more likely to have a particular shape. This may also have some consequences for the evolution of recombination.

Using a certain subset of the non-standard tests that the authors called double-double (d-d) tests (since they compared double mutants with each other), the authors showed that the distribution of epistasis for these d-d tests was slightly more skewed towards positive values than the corresponding distribution of standard tests (figure 4.1). From this observation and from the previous empirical evidence of the existence of compensatory mutations, the authors predicted that the "extent of compensation" or in other words, the strength of positive epistasis must be proportional to the deleterious effect of the mutations. In order to test this prediction, the authors plotted the strength of epistasis (i.e. the standard and non-standard d-d tests e.g e=ar · bs-as · br) against the mean deleterious effect of mutations (e.g. Δ=bs-(as+br)/2, where a, r, b and s label the fitness of distinct single mutations, the fittest genotype comes with a positive sign because we are looking for positive epistasis between deleterious mutations) and also accounted for the statistical dependence of the two. As a result, they found a marginally significant correlation (one-tailed p value was 0.105 for the standard tests and 0.056 for the non-standard tests) between epistasis and mutational effect. The fact that the correlation was more significant for the non-standard tests renders them indispensable.

Some recent studies have also found support for a model of antagonistic epistasis between beneficial mutations and this epistatic interaction is larger for mutations with larger benefit. This phenomena is called diminishing returns epistasis. Interestingly, Fisher's geometric model also predicts a similar pattern of epistasis [37].

Actually, both diminishing returns epistasis and compensatory mutations hypothesis seem to be intuitively clear when one assumes that the combined effect of mutations (whether deleterious or beneficial) must saturate at some point. That is, if two highly beneficial mutations combine, their net effect should be less than the sum of the independent contributions (negative epistatic interaction) and this suppression of the effect must increase as the effect of the mutations increases. Under the saturation assumption, same should hold for the effect of deleterious mutations, only in this case there would be positive epistasis.

Taking a cue from [36], I decided to test the diminishing returns hypothesis using shape theory for other empirical landscapes. Since they had already considered a landscape with deleterious mutations, I chose three 4 locus landscapes, one comprising of synonymous mutations[1] [38], one of small effect beneficial mutations and one of large effect beneficial mutations [39]. I discuss my results in the next section.

4.2 New results

The landscapes that I considered are shown in figures 4.2 and 4.3. The mutations were in the antibiotic resistance enzyme TEM-1 $\beta-$lactamase. This enzyme provides resistance by inactivating penicillin and cephalosporin antibiotics by hydrolyzing their beta-lactam ring. However, further mutations in this gene allows it to attack an extended spectrum of β-lactam antibiotics like cefotaxime (Ctx) [38]. In fact, the resistance to Ctx was taken as a proxy for fitness.

I generated the Markov basis for 4-loci landscapes using the computer algebra system Macaulay2[2]. It comprises of 55 $(= 24 + 24 + 7)$ elements or epistasis tests. Further, the landscapes I considered weren't singly peaked like Elena and Lenski's, rather they were quite rugged. Thus, it didn't make sense to distinguish standard tests from non-standard ones, as was the case in

[1] Synonymous mutations arise due to the redundancy of the genetic code leading to multiple codons corresponding to the same amino acid. They were originally thought to make no difference on the phenotypic level but this was proved otherwise.

[2] http://www.math.uiuc.edu/Macaulay2/

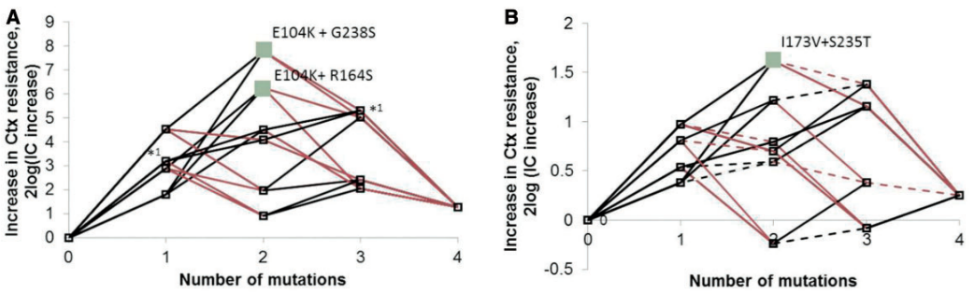


Figure 4.2: The large and small effect $\beta-$lactamase landscapes. Source: [39]

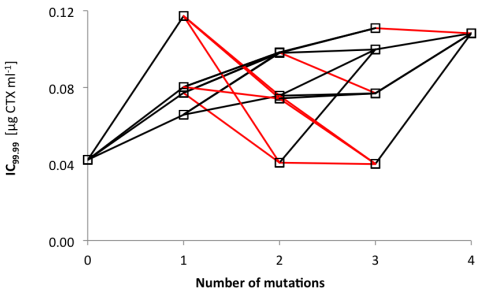


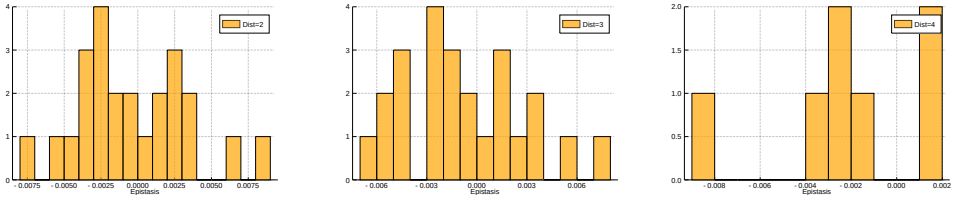
Figure 4.3: The synonymous effect $\beta-$lactamase landscape. Source: [38]

the paper. Therefore, I considered all the tests as measures of the strength of epistasis. These tests span distances 2, 3 and 4 in the Hamming space. There are 24 distance 2 and distance 3 tests, while only 7 distance 4 tests. For the sake of consistency, I placed the monomial with the genotype farthest from the wild type (and consequently also the genotype nearest to the wild type) on the left-hand side of the negative sign (as is typically the case in epistasis tests).

The epistasis distributions for the three landscapes are shown in figures 4.4, 4.5 and 4.6.

For the synonymous landscape, most of the tests, regardless of the distance they cover, show negative epistasis. For the record, 14/24 distance 2 tests, 15/24 distance 3 tests and 5/7 distance 4 tests exhibit negative epistasis. This is also evident from the abundance of the negative epistasis motif in the higher dimensional structure of the fitness landscape which is shown in figure 4.3. That said, the strength of the episatsis (e) per se is not very strong and $|e| < 0.009$.

The small effect landscape shows nearly the same strength of epistasis as the synonymous landscape but the distribution is more strongly skewed to-

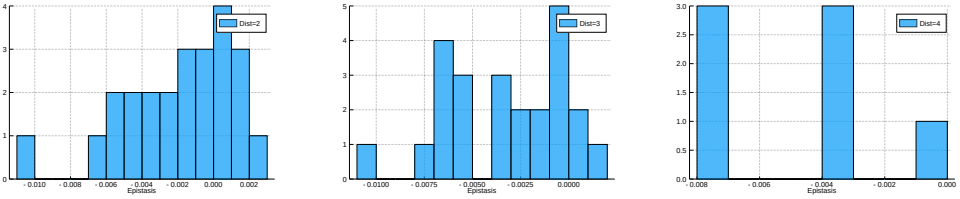


(a) Distribution of distance 2 tests

(b) Distribution of distance 3 tests

(c) Distribution of distance 4 tests

Figure 4.4: Distributions of epistasis for the synonymous landscape.



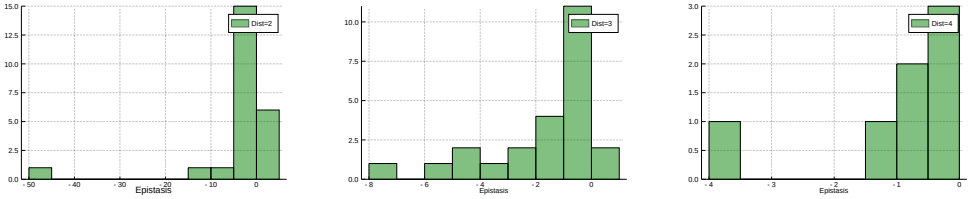
(a) Distribution of distance 2 tests

(b) Distribution of distance 3 tests

(c) Distribution of distance 4 tests

Figure 4.5: Distributions of epistasis for the small effect landscape.

wards negative values. Here, 16/24 distance 2 tests, 21/24 distance 3 tests and 7/7 distance 4 tests show negative epistasis. This pattern is expected in light of the diminishing returns hypothesis because all the individual mutations have beneficial effects and so we expect them to interact antagonistically.



(a) Distribution of distance 2 tests

(b) Distribution of distance 3 tests

(c) Distribution of distance 4 tests

Figure 4.6: Distributions of epistasis for the large effect landscape.

In contrast, the large effect landscape shows much higher strengths of epistasis. Moreover, the distributions are even more skewed towards negative values: 18/24 distance 2 tests, 22/24 distance 3 tests and 7/7 distance 4 tests show negative epistasis. In fact, the distance 2 test $w_{0001} \cdot w_{0111} - w_{0011} \cdot w_{0101}$ shows an exceptionally high value of epistasis. This test and the concerned

genotypes are identified in figure 4.7. The test is between the second and the third loci, with the first and fourth loci held fixed at 0 and 1 respectively. Very strong reciprocal sign epistasis is observed, which is not surprising because the intermediate genotypes (1010 and 1100) are local maxima. The biological reason for this exceptionally high epistasis ought to be investigated further.

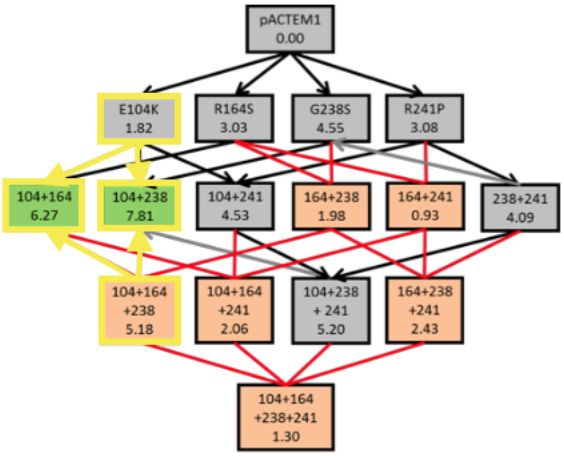
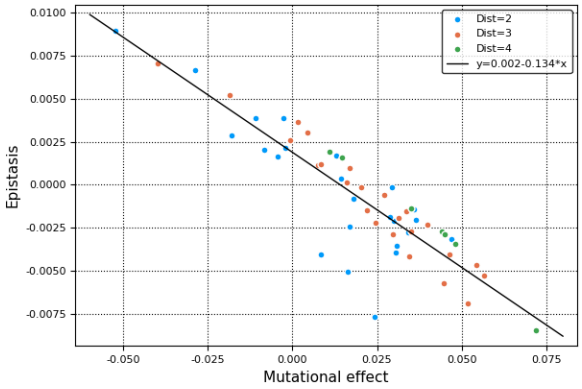


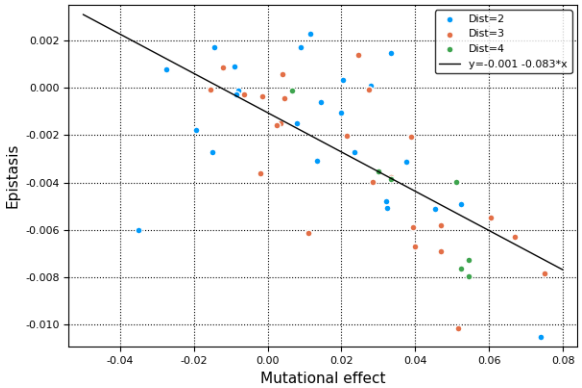
Figure 4.7: The strong epistatic interaction motif is highlighted in yellow in the fitness graph of the large effect landscape that was presented in **[39]**.

All these epistasis distributions point towards a correlation between the strength of negative epistasis and the size of the beneficial effect of the mutations. This is investigated in figure 4.8, where the epistasis strength is plotted against the mutational effect. In the plot for the synonymous landscape, a clear correlation between epistasis and mutational effect can be seen. When the mutations have deleterious effects, the correlation is in agreement with the trend of compensatory mutations, while when the effects are beneficial, the diminishing returns epistasis trend can be observed. A similar correlation is observed for the weak effect landscape, although the values are more scattered. The large effect landscape also appears to roughly follow this trend, although the epistasis seems to be constrained to only negative values, even when the mutational effect is negative.

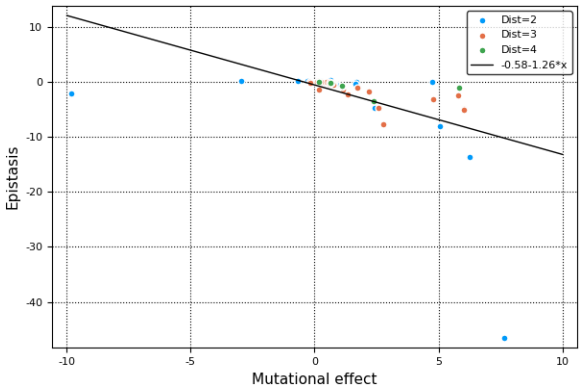
Finally, let's look at the shapes of these three fitness landscapes. The vertices of the simplices in the triangulation of the landscapes are shown in figure 4.9. In each case, the 4-D hypercube is triangulated into 24 simplices. The shape of the large effect landscape is a good example of how the shape informs us about the underlying interactions. All the simplices in the triangulation contain the genotype 1010 (highlighted in red), which makes sense



(a) The synonymous landscape



(b) The small effect landscape



(c) The large effect landscape

Figure 4.8: Epistatic strength versus mutational effect for all the three landscapes.

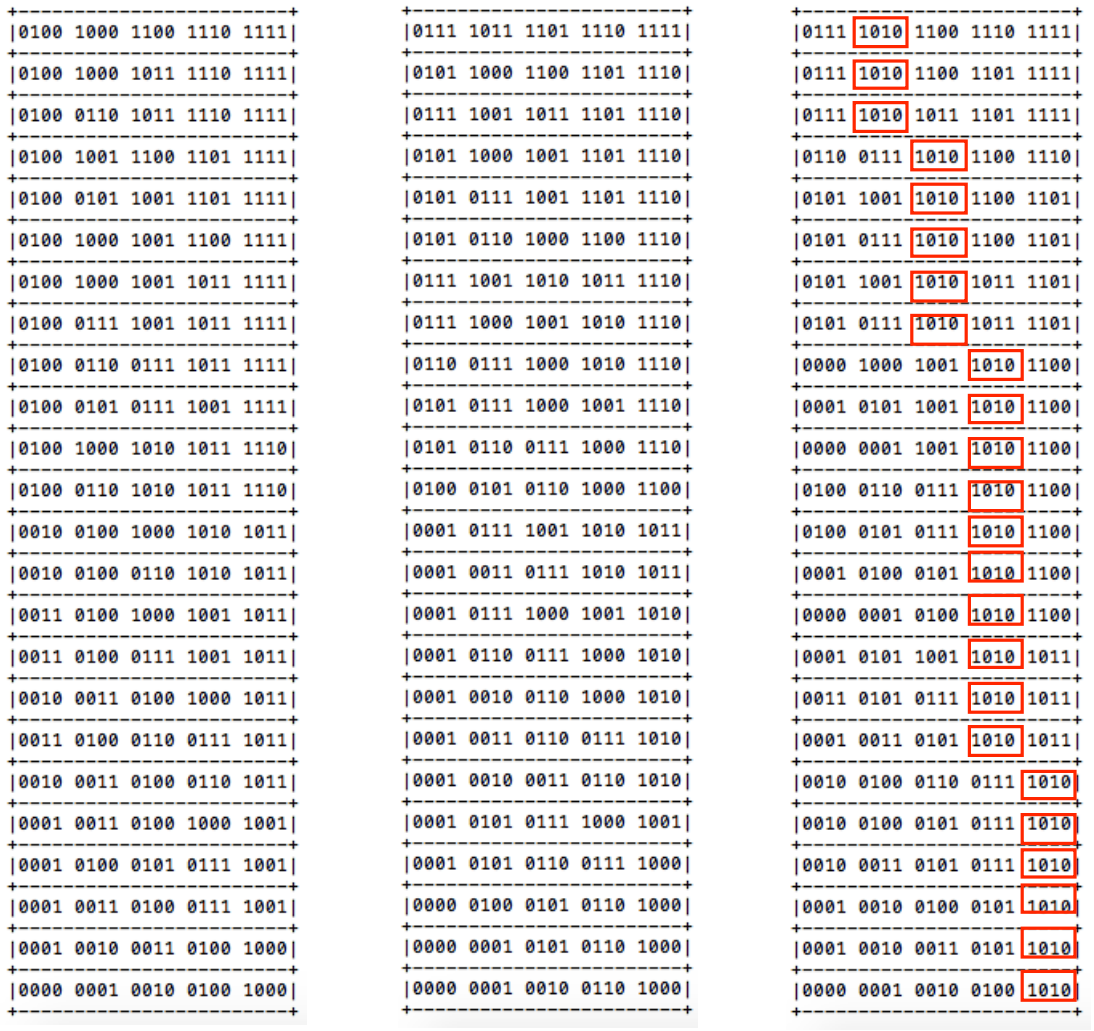


Figure 4.9: The vertices of the simplices in the triangulation of the synonymous, small effect and large effect landscapes (left to right).

given that it is the fittest genotype and is also involved in the strongest pairwise interaction. Similar, but not so striking, is the recurrence of the genotype 1111 in the triangulation of the synonymous landscape and of the genotype 1110 in that of the small effect landscape. This indicates that maximally fit

populations are very likely to contain these genotypes. Moreover, the genotype 0000 only occurs as the vertex of one simplex in the triangulation of the synonymous landscape. This means that it is "sliced off" or very unlikely to occur in maximally fit populations. The genotype 1111 is similarly "sliced off" in the triangulation of the small effect landscape. This makes intuitive sense because as can be seen from the landscapes in figures 4.2 and 4.3, the sliced off genotypes have very low fitness.

It is indeed nice to see how merely the shape of the landscape can tell us something about the outcome of evolution on the landscape.

Chapter 5

Shapes and evolution: Mutation-Selection

In the next chapters, the hypothesis that the shape of a fitness landscape determines the evolutionary trajectory is tested. We expect this hypothesis to be true because we know that epistasis plays a crucial role in evolution and shapes are simply a way of summarising epistatic interactions. This chapter focuses on the simpler case of deterministic evolution with only mutation and selection. The effect of recombination is included in the next chapter.

5.1 Mutation-selection dynamics

The recursion relation for mutation-selection dynamics in discrete time is as follows:

$$x_i(t+1) = \sum_j \mu^{d(\sigma_i,\sigma_j)} \cdot (1-\mu)^{L-d(\sigma_i,\sigma_j)} \frac{w_j}{\bar{w}(\vec{x},t)} x_j(t) \tag{5.1}$$

where, $\vec{x}$ is the vector of genotype frequencies, x_i is the frequency of the ith genotype, μ is the mutation probability, L is the length of the sequences and $d(\sigma_i,\sigma_j)$ is the Hamming distance between the sequences σ_i and σ_j, w_i is the fitness of the ith genotype and $\bar{w}(\vec{x},t)$ is the mean population fitness at time t.

Since the dynamics with only selection and mutation can be linearised **[40]**, many interesting facts about the evolution can be inferred solely from the mutation-selection matrix. Linearising the dynamics proves to be a very powerful step because it considerably reduces the computational time. The linearisation can be done by a transformation of variables which leaves the mutation selection dynamics un-normalised.

Substituting, $z_i(t) = \frac{x_i(t)}{\prod_{\tau=1}^{t-1} \bar{w}(\vec{x},\tau)}$ linearises the evolution and we get:

$$z_i(t+1) = \sum_j \mu^{d(\sigma_i,\sigma_j)} \cdot (1-\mu)^{L-d(\sigma_i,\sigma_j)} w_j z_j(t) \tag{5.2}$$

After working with the un-normalised frequencies $\vec{z}$, one can retrieve the actual frequencies by using the fact that $\vec{x} = \vec{z}/(\sum_i z_i\ )$.

One can also study the dynamics in matrix form, where:

- The mutation matrix M has elements $M_{ij} = \mu^{d(\sigma_i,\sigma_j)} \cdot (1-\mu)^{L-d(\sigma_i,\sigma_j)}$.
- The selection matrix S is a diagonal matrix with $S_{ii} = w_i$ where w_i is the Wrightian fitness of sequence σ_i.
- The dynamics assumes that selection acts first and the census is taken after the mutation step that follows. Thus, the matrix of evolution is $M \cdot S$.

Since all the elements of the transformed mutation-selection matrix are strictly positive, by the Frobenius-Perron theorem **[41]**, the existence of a unique, globally stable equilibrium is guaranteed. The stationary state is given by the eigenvector corresponding to the largest eigenvalue of the matrix (which the theorem guarantees to be real). Further, the rate of convergence to equilibrium is governed by the real part of the second largest eigenvector of the matrix.

Another result is that all eigenvalues of the un-normalised mutation-selection matrix are real. This was shown in **[42]** as follows:

$$M \cdot S\vec{z} = \lambda \vec{z} \tag{5.3}$$

$$\Rightarrow S^{1/2} \cdot M \cdot S\vec{z} = \lambda S^{1/2} \cdot \vec{z} \tag{5.4}$$

$$\Rightarrow S^{1/2} \cdot M \cdot S^{1/2}(S^{1/2} \cdot \vec{z}) = \lambda (S^{1/2} \cdot \vec{z}) \tag{5.5}$$

Let $S^{1/2} \cdot M \cdot S^{1/2} = F$. Then, $F^T = F$ i.e. F is a symmetric matrix with eigenvalue λ and eigenvector $S^{1/2} \cdot \vec{z}$. This implies that all eigenvalues λ of F must be real. Since F and $M \cdot S$ share the same eigenvalues, all eigenvalues of the un-normalised mutation-selection matrix are therefore also real.

5.2 Two locus case

As was mentioned before, shapes are almost trivial for the 2 locus case. Since there is only one circuit and two shapes of the same type, it is easy to anticipate how the shape will effect the dynamics. However, finding the stationary state of even the two locus mutation selection matrix for general fitness landscapes and mutation probabilities is a complicated problem **[43]**. Therefore, I considered simpler fitness landscapes. Without loss of generality, I assumed the double mutant sequence, 11, to be the fittest genotype with fitness 1 and its antipodal sequence 00, the wild type, to have a fixed fitness of 0.5 (unless mentioned otherwise), so that the range of epistatic variation is symmetric. Also, I considered permutation invariant landscapes, leading to sequences 01 and 10 having the same fitness i.e. $w_{01} = w_{10} = w$ and $0 < w < 1$.

Let us see what the stationary state looks like after making certain assumptions about the mutation probability, μ:

(a) $\mu \to 0$

As the mutation probability becomes negligibly small, the mutation-selection matrix reduces to the diagonal selection matrix because all non-diagonal terms depend on μ and consequently go to zero. In this case, w_{11} is the leading eigenvalue and the corresponding eigenvector is $(0,0,0,1)^{\mathrm{T}}$.

(b) $\mu = 0.5$

For $\mu = 0.5$, $1-\mu = \mu$ and therefore, $\mu(\sigma_j \to \sigma_i) := \mu_{i,j} = \mu^2$. Thus the recursion relation becomes $x_i = (\mu^2 \sum_j x_j w_j)/\bar{w} = \mu^2 = 0.25\ \forall i \Rightarrow$ the leading eigenvalue is $0.25 \cdot \sum_i w_i$ and the corresponding eigenvector is $(0.25, 0.25, 0.25, 0.25)^{\mathrm{T}}$.

(c) $w_{11} \gg w_i\ \forall i \neq 11$

In the case of a singly, highly peaked landscape (assumed WLOG to be peaked at 11), all terms of the type $w_i/w_{11} \to 0\ \forall i \neq 11$. Under this assumption,

$$x_{11}^{\mathrm{eq}} = (\sum_j \mu_{11,j} w_j x_j^{\mathrm{eq}})/\bar{w}$$

$$= w_{11}(\sum_j \mu_{11,j}(w_j/w_{11}) x_j^{\mathrm{eq}})/\bar{w}$$

$$\approx (w_{11}\mu_{11,11}x_{11}^{\mathrm{eq}})/(x_{11}^{\mathrm{eq}} w_{11}) = \mu_{11,11}$$

$$= (1-\mu)^2$$

Similarly,

$$x_{10}^{eq}, x_{01}^{eq} = -\frac{(1-\mu)^2}{2w_i} + \frac{(1-\mu)}{2}\sqrt{\frac{(1-\mu)^2}{w_i^2} + \frac{4\mu(1-\mu)}{w_i}} \tag{5.6}$$

where $i = 01$ or 10 respectively and

$$x_{00}^{eq} = -\frac{(1-\mu)^2}{2w_{00}} + \frac{(1-\mu)}{2}\sqrt{\frac{(1-\mu)^2}{w_{00}^2} + \frac{4\mu^2}{w_{00}}} \tag{5.7}$$

These expressions can be further simplified for small w_i and μ. An expansion of the square root till second order leads to :

$$x_i^{eq} \approx \mu(1-\mu(1+w_i)) \text{ for } i = 01, 10 \tag{5.8}$$

and

$$x_{00}^{eq} \approx \mu^2 - \frac{\mu^4 w_{00}}{(1-\mu)^2}. \tag{5.9}$$

After the brief introduction to mutation-selection dynamics, I will now discuss some results.

Figure 5.1 shows how the shape affects the mean fitness at equilibrium. By "shape" here I mean the magnitude of the circuit $e = w_{00} \cdot w_{11} - w_{10} \cdot w_{01}$, even though all landscapes with $e > 0$ have one shape and those with $e < 0$ have the other shape. Obviously, this comparison only makes sense when the shape is varied minimally, i.e. the height or the location of peaks in the fitness landscape are not changed significantly. The variation is only enough to change the value of the epistasis e.

The results in figure 5.1 are not surprising. Given the assumptions, all selection does is to drive the entire population to the fittest genotype while mutation tries to hinder this adaptation by spreading the population away from the peak. Now, for shapes with $e < 0$, the single mutants are quite fit themselves and thus the spreading of the population by mutation doesn't cost too much. Similarly, $e > 0$ implies that the single mutants have a low fitness and thus population spreading becomes very costly, causing a significant decline in the mean equilibrium fitness for large mutation probabilities. Thus, under my assumptions, populations evolving on shapes with negative epistasis have larger equilibrium mean fitnesses than those with positive epistasis. Further, this effect is obviously more prominent for relatively larger mutation probabilities because for small mutation probabilities, selection is strong

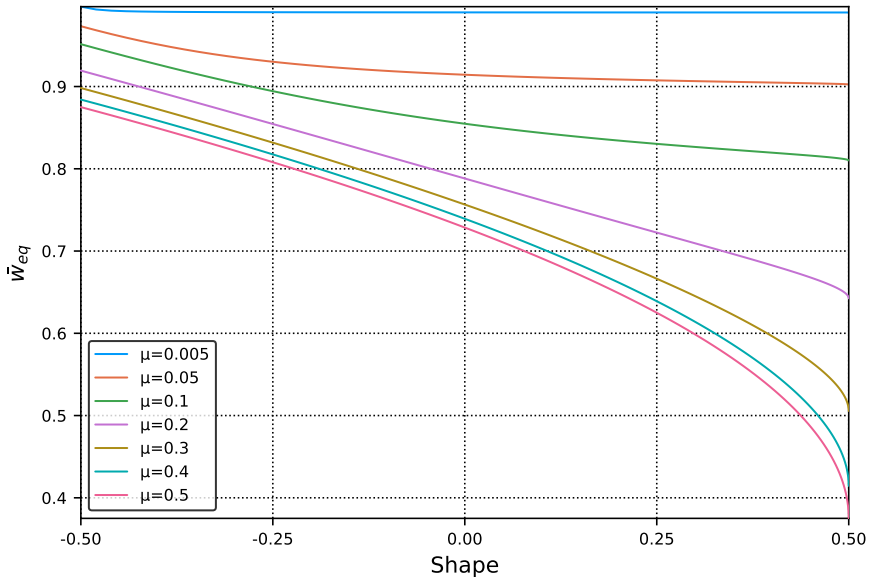


Figure 5.1: Variation of the equilibrium mean fitness with the shape of the permutation invariant landscape for different values of μ.

enough to drive the entire population to the peak and hence the shape (more precisely the value of e) has no effect on the equilibrium mean fitness.

However, a larger equilibrium mean fitness does not in the least imply ease of evolution, i.e. the relaxation time to equilibrium. This is the quantity that I investigated in figure 5.2.

The equilibration time (T_{eq}) can be estimated from the linearised dynamics as follows:

$$\vec{z}(t) = (M \cdot S)^t \cdot \vec{z}(0)$$

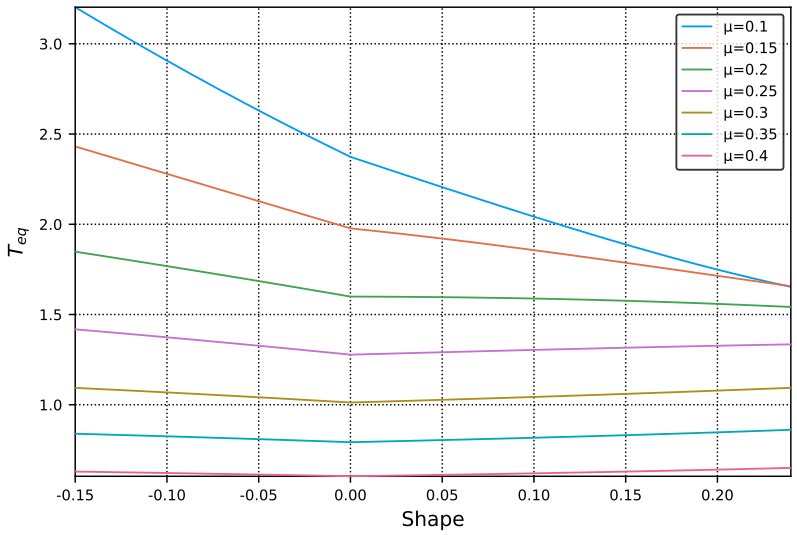
The initial un-normalised frequency vector $\vec{z}(0)$ can be written as a linear combination of the eigenvectors of the matrix $M \cdot S$ i.e. $\vec{z}(0) = \sum_i c_i \vec{v}_i$, where c_i represents the contribution of the ith eigenvector.

$$\Rightarrow (M \cdot S)^t \vec{z}(0) = \sum_i c_i \lambda_i^t \vec{v}_i$$

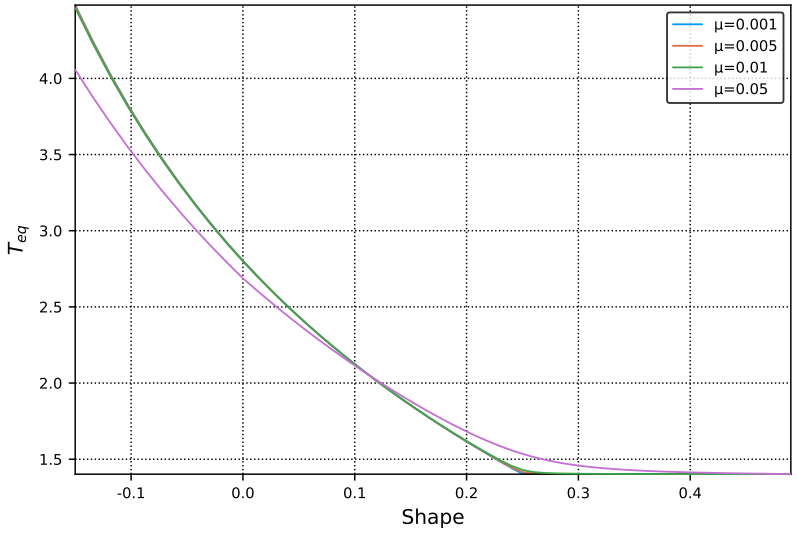
Let λ_1 be the largest eigenvalue, λ_2 the second largest and so on. Then:

$$\vec{z}(t) = c_1 \lambda_1^t [\vec{v_1} + \sum_{i=2} (c_i/c_1)(\lambda_i/\lambda_1)^t \vec{v}_i]$$

Now, as $t \to \infty$ the sum above gets dominated by the term containing $(\lambda_2/\lambda_1)^t$. This gives:



(a) Variation of the relaxation time (T_{eq}) with the shape of the permutation invariant landscape for large values of μ



(b) Variation of the relaxation time (T_{eq}) with the shape of the permutation invariant landscape for small values of μ. The sharpness at zero epistasis is not visible in the plot.

Figure 5.2

$$\vec{z}(t) \sim c_1 \lambda_1^t [\vec{v_1} + (c_2/c_1)(\lambda_2/\lambda_1)^t \vec{v_2}]$$

Consequently,

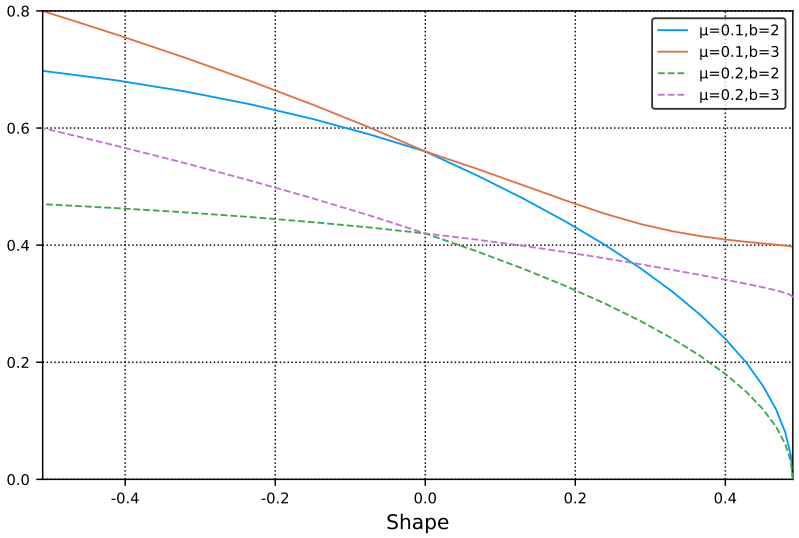
$$x_i(t) = \frac{\lambda_1^i + (c_2/c_1)(\lambda_2/\lambda_1)^t \lambda_2^i}{\sum_j (\lambda_1^j + (c_2/c_1)(\lambda_2/\lambda_1)^t \lambda_2^j)}$$

Here, the superscript on v_1^k represents the kth component of $\vec{v}_1$. This implies the speed of convergence to the equilibrium state is determined by (λ_2/λ_1). The time scale of convergence $(\lambda_2/\lambda_1)^t$ can be expressed as $e^{t \cdot \ln(\lambda_2/\lambda_1)}$, leading to:

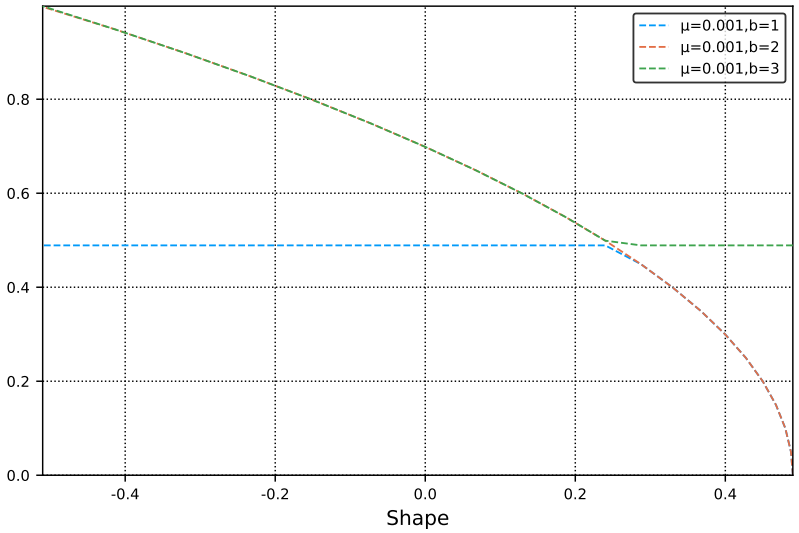
$$T_{\text{eq}} = \frac{1}{\ln(\lambda_1/\lambda_2)} \tag{5.10}$$

The general trends in figures 5.2 a) and b) are in keeping with expectations: T_{eq} expectedly decreases as μ increases and it is larger for negatively epistatic landscapes, which makes sense because in negatively epistatic landscapes, the single mutants are themselves quite competent and this makes it harder for the population to concentrate on the fitness peak. However, the variation of T_{eq} with the shape is not smooth and incidentally, a sharpness in the curve occurs exactly at the transition point between the two shapes, i.e. when the epistasis becomes zero. For small mutation probabilities, this corner in the curve appears at two places- one at zero and one at some positive value of epistasis. For very small mutation probability, $\mu < 0.01$, the curves nearly coincide.

The pointedness of the curves can be explained and understood by looking at the eigenvalues and eigenvectors of the un-normalised matrix $M \cdot S$. T_{eq} depends on the ratio of the largest and second largest eigenvalues, λ_1/λ_2. While λ_1 varies smoothly with the shape, the variation of λ_2 exhibits this non-differentiability exactly at zero epistasis. This is shown in figure 5.3 a). The reason behind this non-differentiability is that the curves of the second and third largest eigenvalues cross at zero epistasis, causing them to interchange their ranks. The implication of this non-differentiability for the population dynamics is as follows: Negatively epistatic landscapes have relatively fitter single mutants, so in this regime, the state corresponding to the single mutants being abundant in the population most strongly competes with the actual stationary state. This also means that this state contributes most to the relaxation time of the dynamics. As one transitions to the positively epistatic landscapes, the single mutants have lower fitnesses and the main contender of the stationary state becomes the state with the population concentrated at



(a) Variations of λ_2 (b=3) and λ_3 (b=2) with the shape for large values of μ.



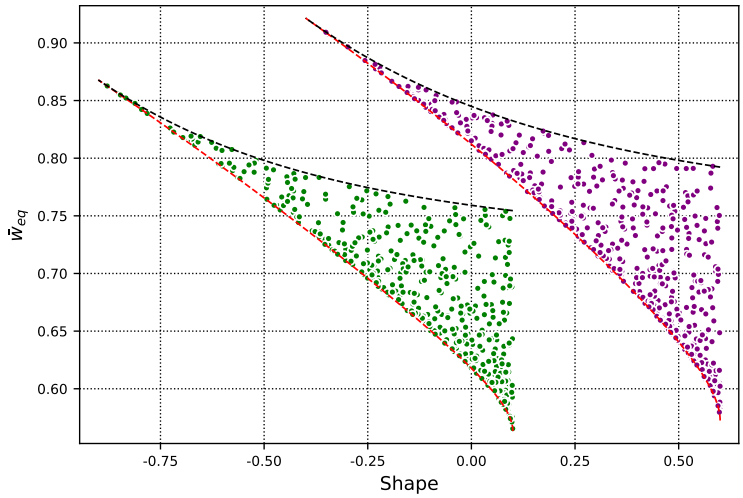
(b) Variations of λ_2 (b=3),λ_3 (b=2) and λ_4 (b=1) with the shape for a small value of μ.

Figure 5.3

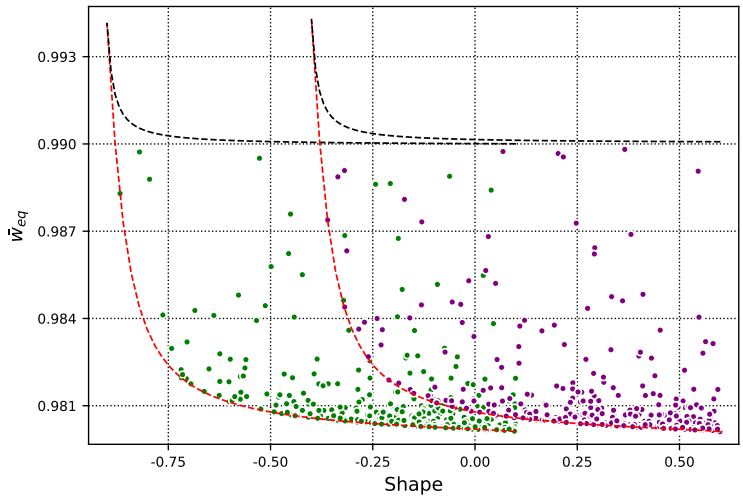
the wild type sequence and contributes most to the relaxation time. It is basically this change in the primary contender of the actual stationary state that is reflected in the non-differentiability of the T_{eq} curves.

A similar argument suffices to explain what we see in figure 5.3 b) for small mutation probabilities. However, in this case, all three eigenvalues apart from the largest one eventually interchange their ranks. For negatively epistatic landscapes, the strongest competing state is the one in which the population is concentrated on either of the single mutants. As we cross zero and move to more positive values of epistasis, this state becomes the one in which the population is equally distributed amongst the two single mutants. Finally, for larger values of positive epistasis, this state represents one in which the population is concentrated on the wild type sequence. This can be better understood in terms of the mutation-selection matrix. For very small μ, $M \sim \mathbb{1}$, where $\mathbb{1}$ is the identity matrix. $\Rightarrow M \cdot S \sim S$. The eigenvalues of S are just the fitness values. Thus, the largest eigenvalue is $\lambda_{11} = w_{11} = 1$ and the corresponding state is the population concentrated on the double mutant. Since, I'm using permutation invariant landscapes, $w_{01} = w_{10} \Rightarrow \lambda_{22} \sim \lambda_{33}$, but they are not equal. Near zero epistasis, the eigenvalues become equal and then they switch their order. For positive enough epistasis, $w_{01} = w_{10} < w_{00}$ and this explains the second crossing and the corresponding change in the strongest competing state.

Next, I decided to relax the constraint of permutation invariance of fitnesses. In that case, for fixed values of epistasis and the fitnesses w_{00} and w_{11}, one gets a range of possible equilibrium mean fitnesses ($\bar{w}_{\text{eq}}$). This is depicted in figure 5.4 for two different mutation probabilities. The bounds on $\bar{w}_{\text{eq}}$ can be found by maximising and minimising it w.r.t w_{01} and w_{10} under the constraint that $w_{10} \cdot w_{01} = w_{00} \cdot w_{11} - e$. It turns out that $\bar{w}_{\text{eq}}$ is bounded from below by permutation invariant landscapes and from above by landscapes with maximally distant fitness values, i.e. $w_{01} = 1$ and $w_{10} = w_{00} \cdot w_{11} - e$ or vice versa. Obviously, for smaller mutation probabilities, the range of variation for a given value of epistasis is much smaller, i.e. $\max(\Delta\bar{w}_{\text{eq}}) \sim 10^{-2} \sim \mu$ (figure 5.4 b)), as opposed to $\max(\Delta\bar{w}_{\text{eq}}) \sim 0.3 \sim \mu$ (figure 5.4 a)) for larger mutation probabilities. Another observation is that $\Delta\bar{w}_{\text{eq}}$ remains nearly constant for all values of epistasis for small μ, while it broadens for large μ. These results indicate that for small μ, the shape doesn't really affect the stationary state of the dynamics because selection is strong enough to attain its motive of pulling the population to the peak so at equilibrium, the rest of the landscape becomes irrelevant. The broadening of the envelope for large μ and positive values of epistasis indicates that the shape constrains the equilibrium state of the dynamics more strongly for negatively epistatic landscapes as opposed



(a) $\mu = 0.25$; green dots represent landscapes with $w_{00} = 0.1, w_{11} = 1$ while purple dots represent those with $w_{00} = 0.6, w_{11} = 1$



(b) $\mu = 0.01$; green dots represent landscapes with $w_{00} = 0.1, w_{11} = 1$ while purple dots represent those with $w_{00} = 0.6, w_{11} = 1$.

Figure 5.4: The bounds on mean equilibrium fitness of landscapes of a given shape for two different mutation probabilities. The red line represents permutation invariant landscapes while the black one represents landscapes with maximally distant single mutants. Note that for very large negative epistasis, maximally distant landscapes become permutation invariant, as is evident from the convergence of the red and black lines.

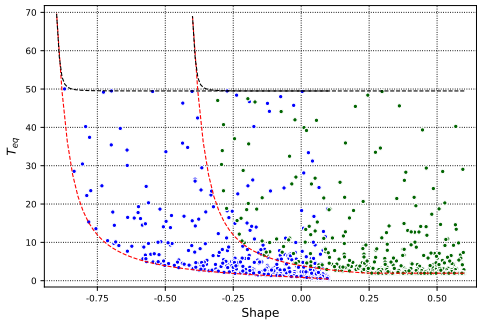
to positively epistatic landscapes.

In figure 5.5, the relaxation time (T_{eq}) appears to be similarly bounded, but the bounds in this case are not as straightforward. On increasing μ from 0.01 to 0.45, we see an interesting trend for the landscapes with $w_{00} = 0.6$: Permutation invariant landscapes (red dashed lines) and landscapes with maximally distant single mutants (black dashed lines) reverse their roles as the lower and upper bound on the range respectively. Already at $\mu = 0.2$, the black line ceases to be the upper bound, then at $\mu = 0.25$, the red and black lines cross each other and they continue to do so as μ is increased. Finally at $\mu = 0.45$, they interchange their roles for positive enough values of epistasis.

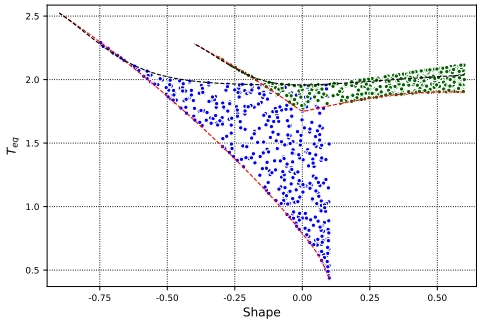
First of all, the average T_{eq} decreases as μ increases, which is what is expected. Further, the range of variation of T_{eq} reduces with increasing μ. This is because the structure of the fitness landscape becomes increasingly less consequential for T_{eq} as μ becomes sufficiently high to overcome most hindrances. Now, the reversal of bounds occurs only for landscapes with a fairly fit wild type sequence (here, $w_{00} = 0.6$) and for high enough mutation probabilities. The reason for this is that for $w_{00} > 0.5$ and positive enough e, the landscape is very likely to have at least one valley. Now, for very small mutation probabilities, we expect the final state to be strongly peaked at 11 and the dynamics can only reach this state by crossing the single mutants. Assuming the dynamics starts with the entire population sitting on the wild type and mutation occurs first and then selection, after the first mutation step, a small fraction of the population will move to the single mutants. Next, depending upon its strength, selection will sweep a fraction of the population uphill, to the sequence 11. Since the strength of selection will be stronger in the permutation invariant case, the stationary state will be reached faster for these landscapes. On the other hand, for large mutation probabilities, we expect the entire population to be uniformly spread over the genotype space. Since selection is strong for permutation invariant landscapes, equilibration takes longer due to the strongly competing forces of mutation and selection. Weak selection along one pathway in landscapes with maximally distant genotypes, enables faster equilibration. This explains the observed trends.

5.3 Three locus case

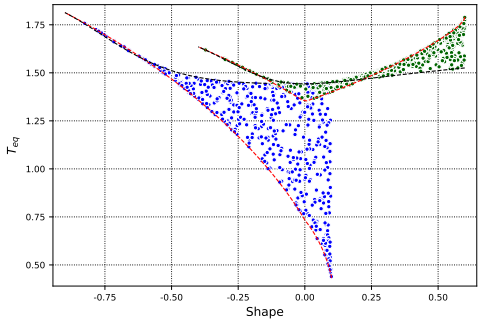
As before, to make sensible conclusions about the dynamics, I fixed the fitnesses of the wild type and its antipodal sequence to 0.1 and 1 respectively. After fixing these values, I generated several realisations of such landscapes, where all the undetermined fitness values were picked from a uniform dis-



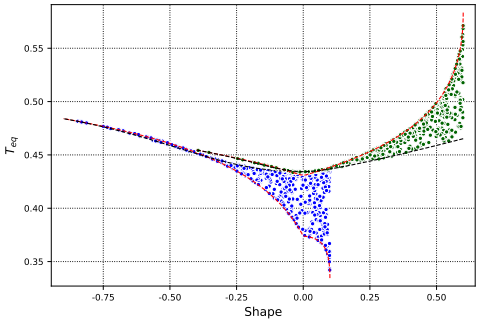
(a) $\mu = 0.01$



(b) $\mu = 0.2$



(c) $\mu = 0.25$



(d) $\mu = 0.45$.

Figure 5.5: The bounds on the relaxation time of landscapes of a given shape for four different mutation probabilities. Blue dots represent landscapes with $w_{00} = 0.1, w_{11} = 1$ while green dots represent those with $w_{00} = 0.6, w_{11} = 1$; red line represents permutation invariant landscapes while black represents landscapes with maximally distant single mutants.

tribution $U(0,1)$. The resultant distribution of shapes, shown in figure 5.6, expectedly looks different from the one observed for HoC landscapes (2.9 a)). Majority of the generated landscapes ($\sim 30\%$) belong to type 1 while only $\sim 4\%$ are of type 6.

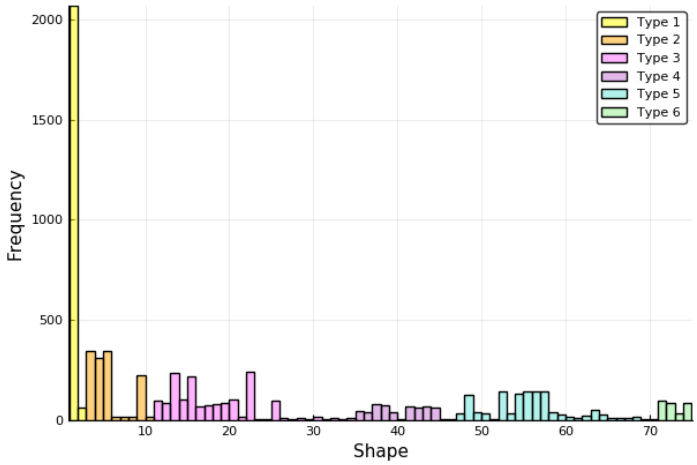
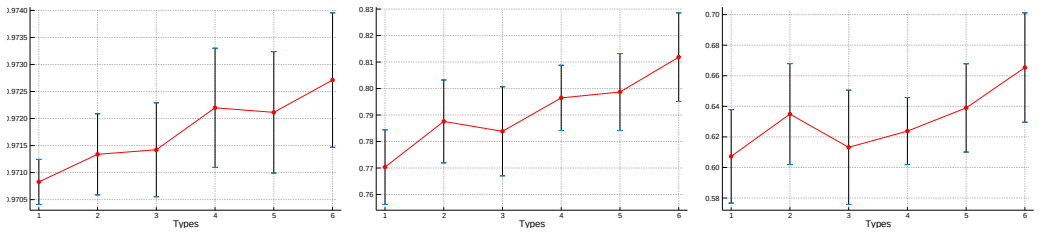


Figure 5.6: The distribution of shapes for landscapes whose wild type and its antipodal sequence have fixed fitness values, here $w_{000} = 0.1$ and $w_{111} = 1$

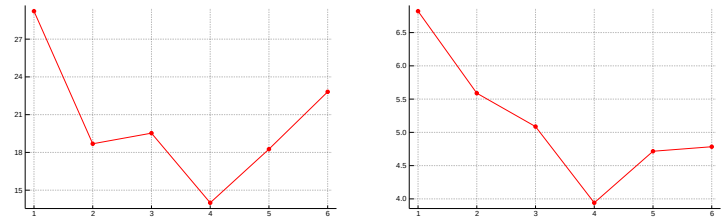
I evolved populations on these restricted landscapes of each type and studied the same two quantities that I looked at in the previous section: Equilibrium mean fitness ($\bar{w}_{\text{eq}}$) and the time to equilibrium (T_{eq}).

The variation of $\bar{w}_{\text{eq}}$ with the type for three mutation probabilities is shown in figure 5.7. At least for small mutation probabilities ($\mu \leq 0.1$), the mean equilibrium fitness increases with the type. While comparing shapes with graphs,



(a) μ=0.01 (b) μ=0.1 (c) μ=0.25

Figure 5.7: Variation of equilibrium mean fitness with the type of the landscape for three mutation probabilities. Error bars represent one half of the standard deviation.



(a) The variation of the means for $\mu = 0.01$

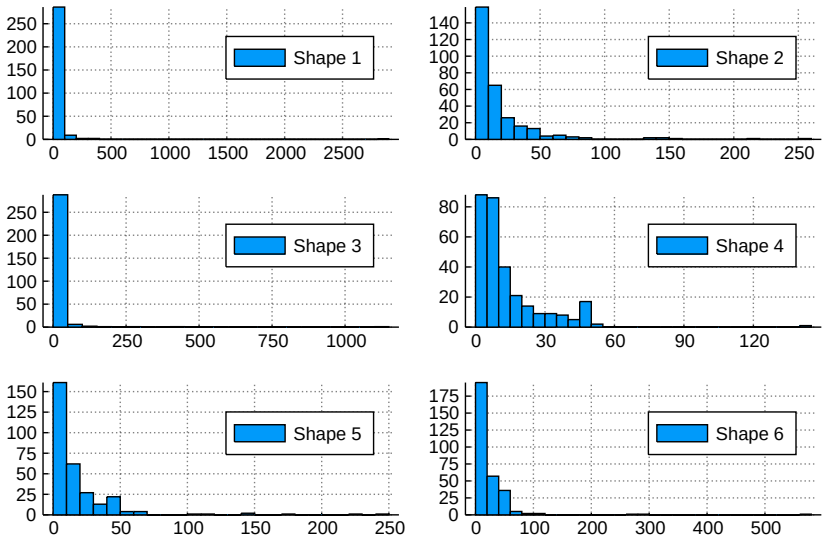
(b) The variation of the means for $\mu = 0.1$.

Figure 5.8: Variation of mean of the relaxation time (T_{eq}) with the type.

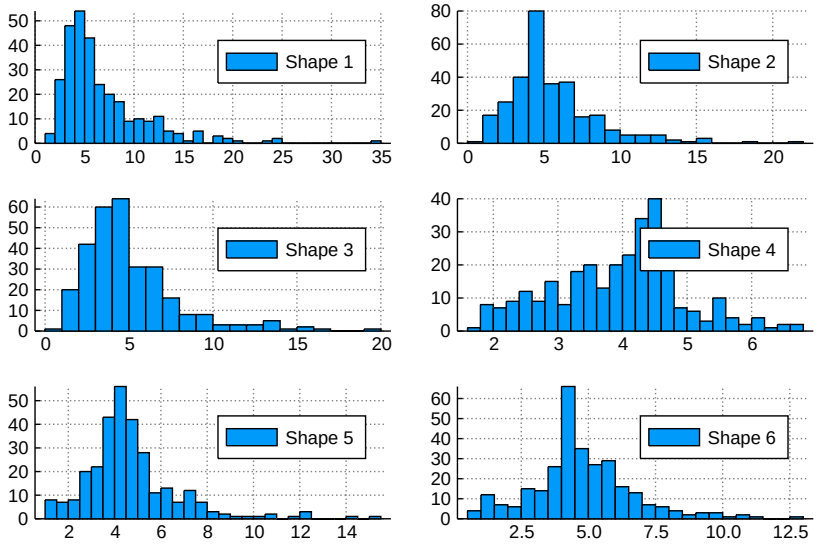
I found that type 1 landscapes, on average, have maximum ruggedness (or number of peaks) and type 6 landscapes have the least (see figure 3.1). This variation in ruggedness over the types can explain the observed variation of the equilibrium mean fitness because having a small mutation probability means being below the mutational threshold and thus single peaked landscapes fare better than relatively "flat" ones with multiple peaks. As the mutation probability increases beyond 0.1, the monotonic increase becomes less pronounced. However, one must not overlook the fact that for small mutation probabilities, where this monotonic increase is more apparent, the magnitude of increase is very small. The variation in the equilibrium mean fitness increases with increasing the mutation probability. This signifies that the equilibrium mean fitness becomes more sensitive to the type of the landscape for higher mutation probabilities.

Figure 5.8 shows the variation of the mean of T_{eq} for two mutation probabilities. Naturally, T_{eq} is larger for smaller mutation probabilities. Moreover, for $\mu = 0.01$, no striking pattern can be inferred due to the large variance of the distributions, while for $\mu = 0.1$, a clearer trend emerges wherein the mean of T_{eq} decreases with the type. It is more interesting to note how different the distributions for each type look for the two mutation probabilities (figure 5.9). For small μ, the distribution is very heavy tailed, while for large μ, it is more symmetric and lighter tailed.

A similar feature observed for both the quantities is that the intra type variance changes significantly with μ. While in the case of $\bar{w}_{\text{eq}}$, it increases with increasing μ, for T_{eq}, it decreases with increasing μ. This is in keeping with one's intuition. Despite using a restricted set of landscapes, the wide variance about the mean values of $\bar{w}_{\text{eq}}$ and T_{eq} highlights that the types do not constrain the dynamics so much.



(a) The distribution per shape for $\mu = 0.01$



(b) The distribution per shape for $\mu = 0.1$

Figure 5.9: Distributions of the relaxation time (T_{eq}) for the different types.

Chapter 6

Shapes and evolution: Recombination

In this chapter, an additional step of recombination has also been included in the dynamics. The primary focus will be on the effect of shape on the evolution of recombination. This study is motivated by the fact that for 2 locus landscapes, recombination is "advantageous" in landscapes of one shape- the shape corresponding to negative epistasis, while it is "disadvantageous" in landscapes of the other shape **[44]**. But before exploring shapes, I will briefly introduce the process of recombination and previous results regarding the question of its evolution.

6.1 Recombination

During Meiosis[1], recombination between non-sister homologous chromosomes[2], leads to the offspring having novel traits that cannot be found in either parent. This is facilitated by chromosomal crossover, which is depicted in figure 6.1.

Further, the probability of recombination between 2 loci on a chromosome depends on the distance between them. Genes that are close to each other are unlikely to produce recombinant gametes, only sufficiently distant ones undergo crossover that can destroy correlations between them. This means, if genes A and b recombine with probability r, then they produce the non

[1] It is a type of cell division that is characteristic of sexual reproduction.

[2] Sister chromatids are formed during cell division and contain the exact same genes and alleles, while non-sister chromatids form homologous pairs and have the same genes but may contain different alleles **[45]**

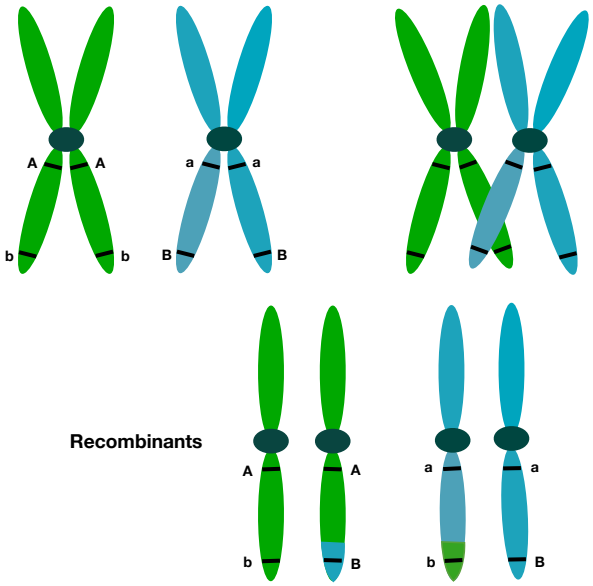


Figure 6.1: The process of chromosomal crossover between homologous chromosomes during Meiosis I. Due to this, the offsprings have a different set of alleles and genes than their parents do. In the diagram, genes B and b are crossed over with each other, making the resulting gametes AB, ab, Ab and aB.

recombinant gametes Ab and aB with probability $(1-r)/2$ and the recombinant gametes ab and AB with probability $r/2$.

If the two genes are as far apart as possible, or even on different chromosomes $r = 0.5$, which is the maximum possible value of r [46]. Genes that have $r < 0.5$ are said to be *linked*. More concretely, linkage refers to the deviation of genotype frequencies from their expected values under the assumption of random association of alleles. Linkage disequilibrium (D) is a measure of how linked two genes are. Mathematically,

$$D := x_{00} \cdot x_{11} - x_{01} \cdot x_{10} \tag{6.1}$$

where x_i represents the frequency of the ith genotype.

Moreover, for a purely recombining population, $D(t)$ decays as:

$$D(t) = (1-r)^t D(0) \tag{6.2}$$

Which basically means that recombination breaks down linkage by reducing D by the same fraction in each time step.

In the mathematical models used in this thesis, I have considered uniform crossover between haploid sequences, meaning that when two length L sequences σ and τ recombine, each bit from the offspring's genome is independently chosen from the two parents with equal probability. Then the probability of producing non-recombinant sequences is $(1-r)/2$ each, while the probability of each recombinant sequence is $r/(2^L-2)$.

Thus, the recursion relation for evolution due to recombination is the following:

$$x_i = \sum_{j,k=1}^{2^L} R_{i|jk} x_j x_k \tag{6.3}$$

where $R_{i|jk}$ is the probability that sequences σ_j and σ_k produce an offspring σ_i.

6.2 The evolution of recombination

The prevalence of sexual reproduction in nature is difficult to explain, given its obvious disadvantages– for example, the famous two fold cost of sex in comparison to asexual reproduction, the breaking up of linkage between favourable gene combinations and the excess time, energy and risks associated with sex. Although, a complete explanation for the origin and maintenance of sex still remains elusive, there have been several good attempts.

6.2.1 Direct models

Such models include factors due to which sexual reproduction might have had a direct impact on the mean fitness of the organism and would have had an immediate advantage. One hypothesis that falls under this category is that sex could be a by-product of double strand DNA repair. These models are however difficult to study and test empirically **[47]**. Moreover, they can only explain the origin of sexual reproduction but not so much its maintenance. Therefore, in this thesis, the term 'advantage of sex' will refer primarily to the indirect advantage of sex, which is described in the following section.

6.2.2 Indirect models

Such models are relevant for explaining the maintenance of sex. One way in which sex can have an indirect advantage is by increasing variation (by

breaking linkage between genes) on which directional selection can subsequently act. Doing so either accelerates adaptation or decelerates maladaptation. However, this is only helpful in the presence of negative linkage disequilibrium (D). Negative D means that the double mutant is less common in the population than is expected from random associations of alleles. Given that this is the case, the obvious question to ask is about the origin of this negative D. The most popular answers to this question are: Negative epistasis, genetic drift or a combination of both.

Moreover, several models exist in the literature that try to justify the evolution of recombination by attributing the linkage to different sources. While the classic Fisher-Muller hypothesis (figure 6.2) relies on genetic drift, the Red Queen hypothesis relies on fitness fluctuations over time and the spatial heterogeneity hypothesis on fitness fluctuations over space. **[48]**

Empirical studies have tried to find evidences for each of the possibilities, but it is not yet clear which is the main cause of negative D **[47]**. Since my motive is to study the contribution of epistasis, I will only focus on infinite populations because it was found in **[49]** that for small populations, the contribution of drift outweighs that of epistasis.

6.3 The effect of shapes

Naturally, recombination will evolve and persist in populations only if it confers some advantage. Now there are several ways of quantifying the advantage of recombination and I will be considering the following three:

1. The difference of the mean fitnesses of a population that evolves via mutation, selection and recombination and another that evolves only by mutation and selection, as a function of time.

2. The difference of genotype frequencies of the fittest mutant of a population that evolves via mutation, selection and recombination and another that evolves only by mutation and selection.

3. Spread of a recombination modifier gene that increases the recombination rate.

6.3.1 Two locus case

The effect of shapes in the two locus case is already known. Most of the important conclusions have been drawn by Eshel and Feldman in **[51]**. Firstly,

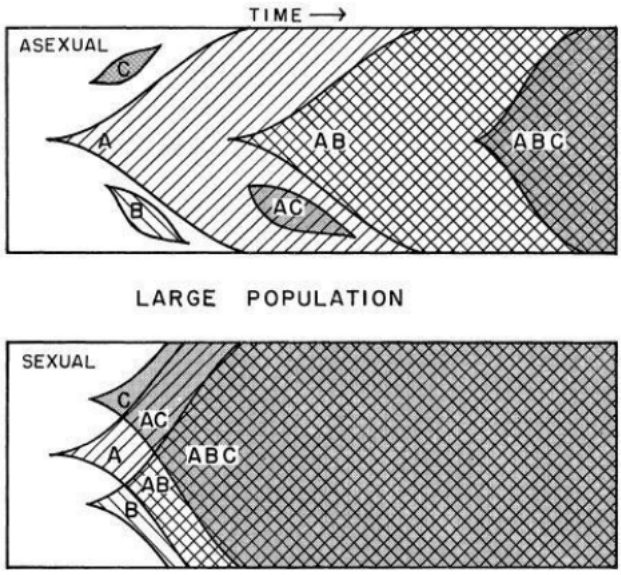


Figure 6.2: Pictorial representation of the Fisher-Muller hypothesis. A, B and C are three individually beneficial mutations that compete with each other due to the finite size of the population. Consequently, they can only sequentially fix in asexual populations. Recombination expedites their incorporation by increasing the abundance of gene combinations. **[50]**

they showed that under deterministic mutation-selection dynamics, for a population with initial $D(t=0)=0$, $D(t)\cdot\epsilon(t)>0$ at all $t>0$, where ϵ represents the epistasis **[44]**. This result can be proved by the principle of mathematical induction. One needs to additionally employ the recursion relations for the change in D due to mutation and selection that were developed in **[52]**. The following is an outline of the proof:

Assuming that $x_{00}(0)=1$

$$\Rightarrow D(0)=x_{00}(0)\cdot x_{11}(0)-x_{01}(0)\cdot x_{10}(0)=0$$

Further, assuming mutation occurs before selection, change in D after the first mutation step,

$$\Delta_\mu D=-4\cdot D(0)\mu(1-\mu)=0$$
$$\Rightarrow D_\mu(1)=0$$

Next, after the first selection step,

$$\begin{aligned} D(1) &= D_\mu(1) + \Delta_s D(1) = \Delta_s D(1) \\ &= p^2(1-p)^2\epsilon/(1+2ps_1+p^2A)^2 \\ &= \alpha\cdot\epsilon \end{aligned}$$

where, $w_{00}=1, w_{01}=w_{10}=1+s_1, w_{11}=1+s_2$, p is the allele frequency for the allele 1, $A=s_2-2s_1$ and $\alpha = p^2(1-p)^2/(1+2ps_1+p^2A)^2 > 0$

$$\Rightarrow D(1)\cdot\epsilon = \alpha\cdot\epsilon^2 > 0$$

This proves that $D\cdot\epsilon > 0$ after the first step of the dynamics. Now, let us assume that this is the case after m steps, i.e. $D(m)\cdot\epsilon > 0 := \beta$. Then, $D(m) = \beta/\epsilon$.

The effect of the mutational step,

$$\begin{aligned} \Delta_\mu D &= -4\cdot D(m)\mu(1-\mu) \\ &= -4\cdot\beta/\epsilon\mu(1-\mu) \Rightarrow D_\mu(m+1) \\ &= \beta/\epsilon - 4\cdot\beta/\epsilon\mu(1-\mu) \end{aligned}$$

and after selection,

$$\begin{aligned} D(m+1) &= x_{00}(m+1)\cdot x_{11}(m+1) - x_{01}(m+1)\cdot x_{10}(m+1) \\ &= (w_{11}w_{00}\cdot x_{00}^{\mu}(m+1)x_{11}^{\mu}(m+1) - w^2(x^{\mu}(m+1))^2)/(\bar{w}^2) \\ &= w^2\cdot D_\mu(m) + x_{00}x_{11}\cdot\epsilon \\ &\Rightarrow D(m+1)\cdot\epsilon = w^2 D_\mu(m+1)\epsilon + x_{00}x_{11}\cdot\epsilon^2 > 0 \ \because D_\mu(m+1)\cdot\epsilon > 0 \end{aligned}$$

Thus, by the principle of mathematical induction,

$$D(t)\cdot\epsilon > 0 \ \forall t \geq 1$$

Further, since the mutation selection equilibrium is independent of the initial condition, $D(T_{eq})\cdot\epsilon > 0$ in general. This is also seen in simulations on random landscapes with particular values of epistasis (figure 6.3).

This result is important because it proves that negative epistasis is a source of negative D in two locus landscapes. This also indicates that negatively epistatic landscapes, by generating negative D, can potentially confer some advantage to recombination. Eshel and Feldman formalised this by proving that for 2 locus, permutation invariant, negatively epistatic fitness landscapes, such that wild type is the least fit and double mutant is the fittest,

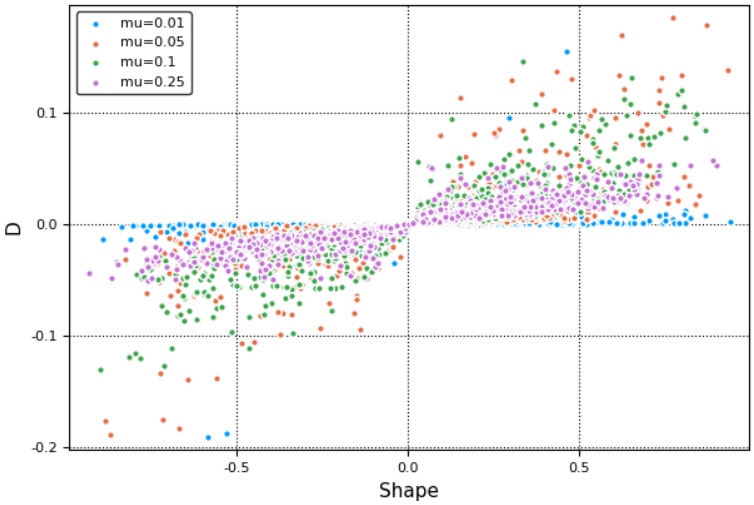


Figure 6.3: Linkage disequilibrium and epistasis (shape) always have the same sign at the mutation-selection equilibrium. Shown here for four different mutation probabilities.

if the dynamics begins with a wild type monomorphic population, at all subsequent times, the double mutant frequency for an asexual population will be smaller than that for a sexual population. The converse is true when epistasis is positive. They thus countered the view that recombination always accelerates evolution. Moreover, they highlighted the importance of the shape of the 2 locus landscape in the evolution of recombination– particularly the fact that recombination is advantageous for negatively epistatic landscapes.

6.3.2 Three locus case

Linkage disequilibirum

As was shown in the previous section, in the two locus case, negative epistasis is a source of negative D. It is also of interest to know if such is the case for 3 locus landscapes. However in the 3 locus case, the problem of measuring D becomes equivalent to the problem of measuring epistasis, i.e. one doesn't know which tests to consider. Thus, one intuitive way to measure D could be to employ the Markov basis of the interaction space here as well. In fact, as shown in figure 6.4, each of the nine elements of the Markov basis that exist for 3 locus landscapes correlates with the corresponding disequilibrium tests. Further, they seem to nearly always satisfy $D_i \cdot e_i > 0$, where $i \in \{1,2,..,9\}$, meaning that the signs of the epistatic interactions in the fitness landscape

guide the signs of D_is at mutation-selection equilibrium. The deviations from this inequality occur only for small values of epistasis. Moreover, the relation between D_i and e_i looks very similar for all i.

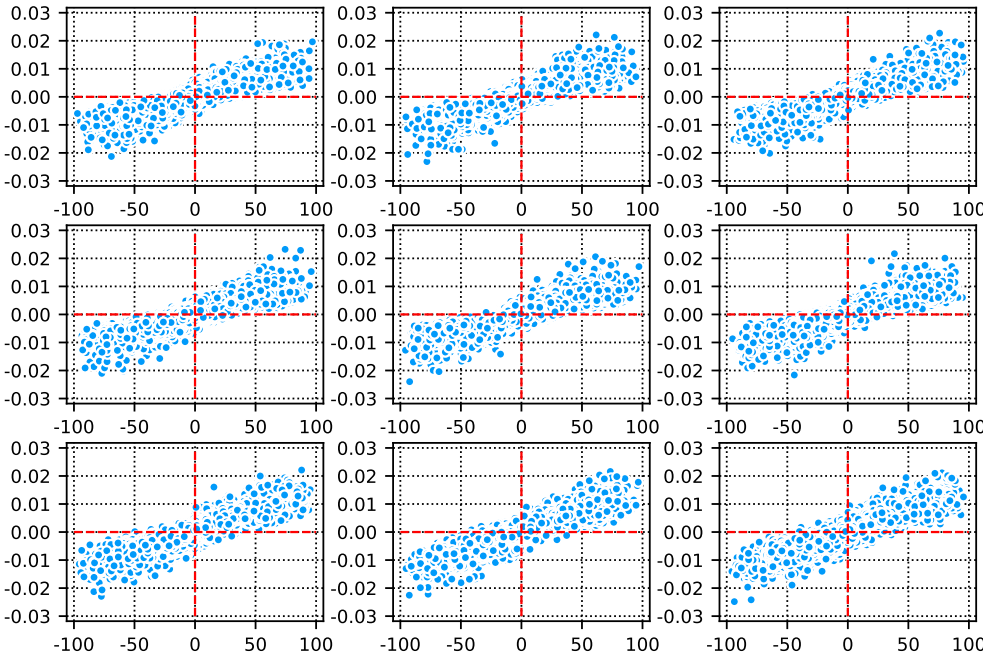


Figure 6.4: Disequilibrium tests (y axis) versus epistasis tests (x axis) for $\mu = 0.25$. $D_i \cdot e_i > 0$ seems to hold for most landscapes.

Permutation invariant (PI) fitness landscapes

It can be inferred from the 3 locus results of the previous chapter, that the shape of a landscape is a relatively weak constraint. Thus, further constraints on the landscape are needed to study characteristic dynamics of populations on particular shapes. Apart from fixing the fitness values of the wild type (to 0.1) and the triple mutant (to 1), I introduced an additional symmetry, namely the permutation invariance of the rest of the genotype to fitness map. This reduces the number of shapes to six and they belong to four types (1,2,4 and 6). The reduced distribution is shown in figure 6.5. It is interesting to note that nearly half of the randomly generated landscapes have shape 1 and none of them belong to type 3 or 5. As a side note, the number of shapes for 4 locus permutation invariant landscapes turned out to be just 41, which is also a significant reduction from the original number of 87,959,448.

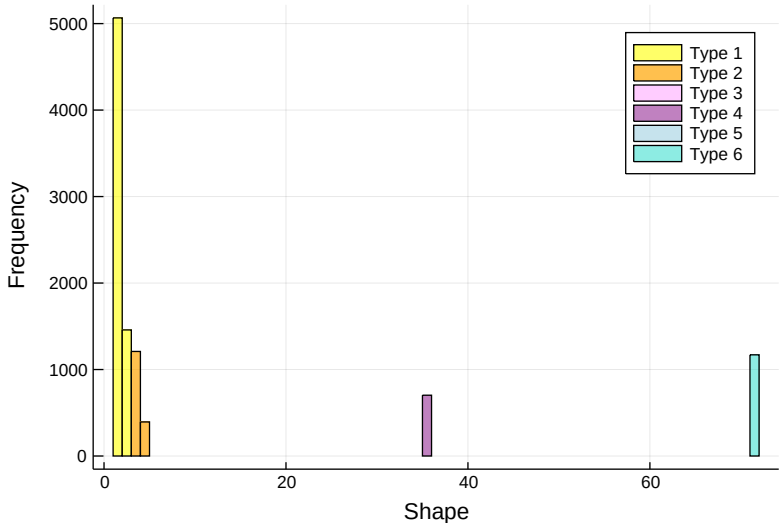


Figure 6.5: Distribution of shapes of permutation invariant landscapes.

It is easy to see why only six possible shapes emerge. Permutation invariance implies $w_{001} = w_{010} = w_{100} := w_1$ and $w_{101} = w_{011} = w_{110} := w_2$. This causes the Markov bases to have only three distinct elements: $a := w_{000} \cdot w_{111} - w_1 \cdot w_2$ which is a measure of the overall epistasis in the landscape, $b := w_{000} \cdot w_2 - w_1^2$ which is a measure of the epistasis between the double mutants and the wild type and $c := w_{111} \cdot w_1 - w_2^2$ which is a measure of the epistasis between the triple mutant and the single mutants. Further, $a = (w_{000} \cdot c + w_2 \cdot b)/w_1$, which means that the sign of a is determined by that of b and c. From simple combinatorics one can infer the possible sign patterns of a, b and c. It is evident that when both b and c have the same sign, a is also compelled to have that sign. This gives rise to 2 shapes. On the other hand, when b and c have different signs, a can have either sign, leading to another four possible sign patterns.

In the remaining part of this chapter, I have relabelled the shapes in figure 6.5 (from left to right) as 6, 4, 5, 2, 3, 1 so that neighbouring shapes differ only by the sign of one element of the Markov basis. This becomes clearer from figure 6.6. Sign patterns of the six shapes are summarised in table 6.1 and some sample landscapes of each shape are plotted in figure 6.7. Moreover, from the characteristic sign patterns of the shapes, it becomes evident that the first four shapes cannot have multiple peaks. This is because having multiple peaks in three locus PI landscapes implies

$$w_1 > w_2 \text{ and } w_1 > w_0 \tag{6.4}$$

where w_i is the fitness of genotypes with i mutations. Shapes 1, 2 and 3 must have

$$\begin{aligned} c = w_3 \cdot w_1 - w_2^2 < 0 \\ \Rightarrow w_1 - w_2^2 < 0 \ (\because w_3 = 1) \\ \Rightarrow w_1 < w_2^2 < w_2 \ (\because w_2 < 1) \end{aligned} \tag{6.5}$$

This means shapes 1, 2 and 3 do not satisfy the requirements of equation 6.4 and hence cannot have multiple peaks. Similarly, shape 4 must have

$$b = w_0 \cdot w_2 - w_1^2 > 0 \tag{6.6}$$

while equation 6.4 $\Rightarrow w_1^2 > w_0 \cdot w_2$. This directly contradicts equation 6.6 and hence shape 4 cannot have multiple peaks either.

Shape	a	b	c
1	<0	<0	<0
2	<0	>0	<0
3	>0	>0	<0
4	>0	>0	>0
5	>0	<0	>0
6	<0	<0	>0

Table 6.1: The sign patterns of the remaining elements of the Markov basis (a, b and c) for the 6 shapes

The reduction in the number of shapes permits an exhaustive analysis of the evolution of recombination on these landscapes. One can already guess the outcome based on the results of Eshel and Feldman **[44]** for the two locus case because the present case is essentially a concatenation of 2 two locus landscapes.

I started by looking at the difference in mean fitness of sexual and asexual populations as a function of time i.e. $\Delta\bar{w}(t) = \bar{w}_{\text{sex.}}(t) - \bar{w}_{\text{asex.}}(t)$. I set my initial condition to be a monomorphic population consisting only of wild type sequences. The results are depicted in figure 6.8. The average over several landscapes is shown in red. In most cases, there is very little difference between the equilibrium mean fitness of the sexual and asexual populations. On the contrary, in the doubly negative landscapes (both $b < 0$ and $c < 0$) of shape 1, a noticeable advantage of recombination persists even at long times. At

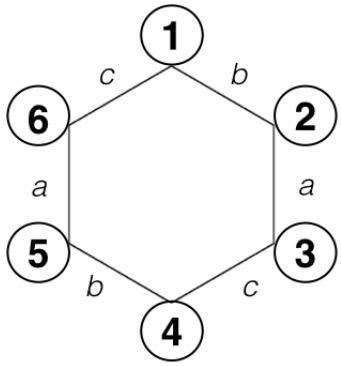
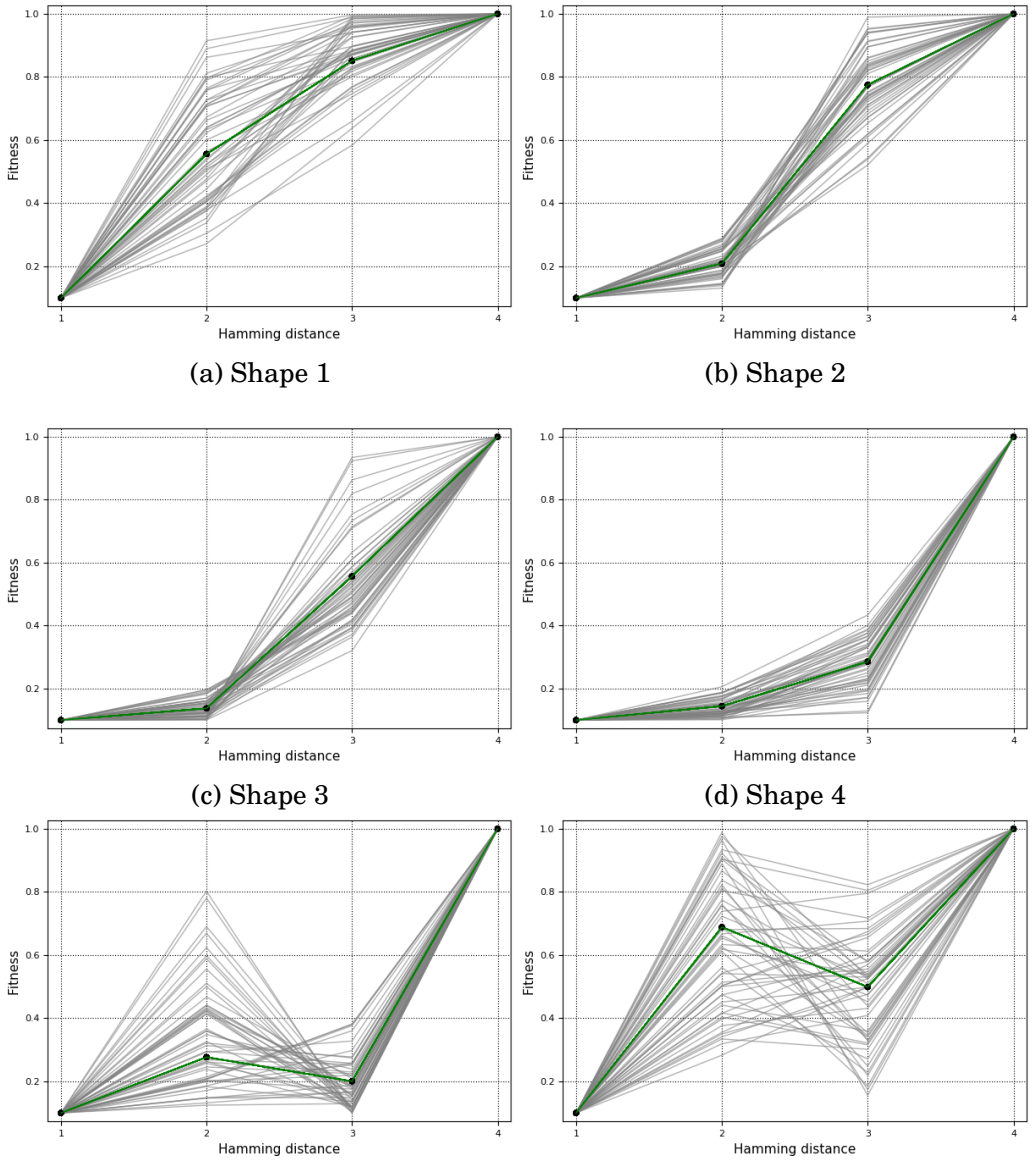


Figure 6.6: The secondary polytope of the shapes of PI landscapes. The labelling on the edges indicate the Markov basis elements in which the neighbouring shapes differ.

the other end of the spectrum is shape 6, where sexual populations on certain landscapes remain significantly maladapted even at equilibrium. This pulls the difference of average mean fitness below zero at long times. Shape 6 also shows the maximum heterogeneity in the response to recombination. This could be because nearly half of the random landscapes have shape 6. The overall behaviour of shapes however is in accordance with what we would expect from Eshel and Feldman's results [44]:

1. The mean fitness of the sexual population remains higher in the beginning for the doubly negative landscapes of shape 1. In that sense, recombination is advantageous for this shape. However, $\Delta\bar{w}(t)$ decreases with time. The advantage is in accordance with what one would expect by extension of the 2 locus result. By a similar extension, one would expect recombination to be disadvantageous for the doubly positive landscapes of shape 4. This is exactly what we see in figure 6.8(d). In fact, $\Delta\bar{w}(t)$ is significantly negative in the beginning, but it very soon reduces to zero.

2. Recombination in both shapes 2 and 3 is initially detrimental due to the positive epistasis in the first part but as the population evolves to see the negatively epistatic second half of the landscape, it begins to benefit from the effect of recombination and this effect then gradually diminishes to zero. This is exactly as we would expect from the extension of the 2 locus results. However, these two shapes differ in the sign of the overall value of epistasis (a). This results in the slightly better performance of recombination in shape 2 landscapes, in terms of the lower



(a) Shape 1

(b) Shape 2

(c) Shape 3

(d) Shape 4

(e) Shape 5

(f) Shape 6

Figure 6.7: Some example landscapes of each shape. The mean of 100 realisations is plotted in green.

average deleterious effect of recombination in the beginning and a more persistent beneficial effect at long times.

3. Shapes 5 and 6 are interesting because they are most likely to have 2 peaks and recombination is known to do a terrible job in multi-peaked landscapes **[53]**. This is also partly what we see in figure 6.8(e) and (f), where in some cases, $\Delta\bar{w}(t)$ is highly negative. However, in shape 5, we do not see any maladapted landscapes at long times, as opposed to shape 6 where some landscapes do not recover from the maladapted state at all.

The dependence of the results on r and μ is as follows: The peak of the $\Delta\bar{w}(t)$ plot for each shape increases with increasing r, while retaining the characteristic form of the curves. The effect of increasing μ is to expedite the process, as a result of which the curves shift to the left. At the same time, the peak $\Delta\bar{w}(t)$ decreases. This is because large values of μ cloud the effect of recombination. These trends for shape 1 are shown in figure 6.9.

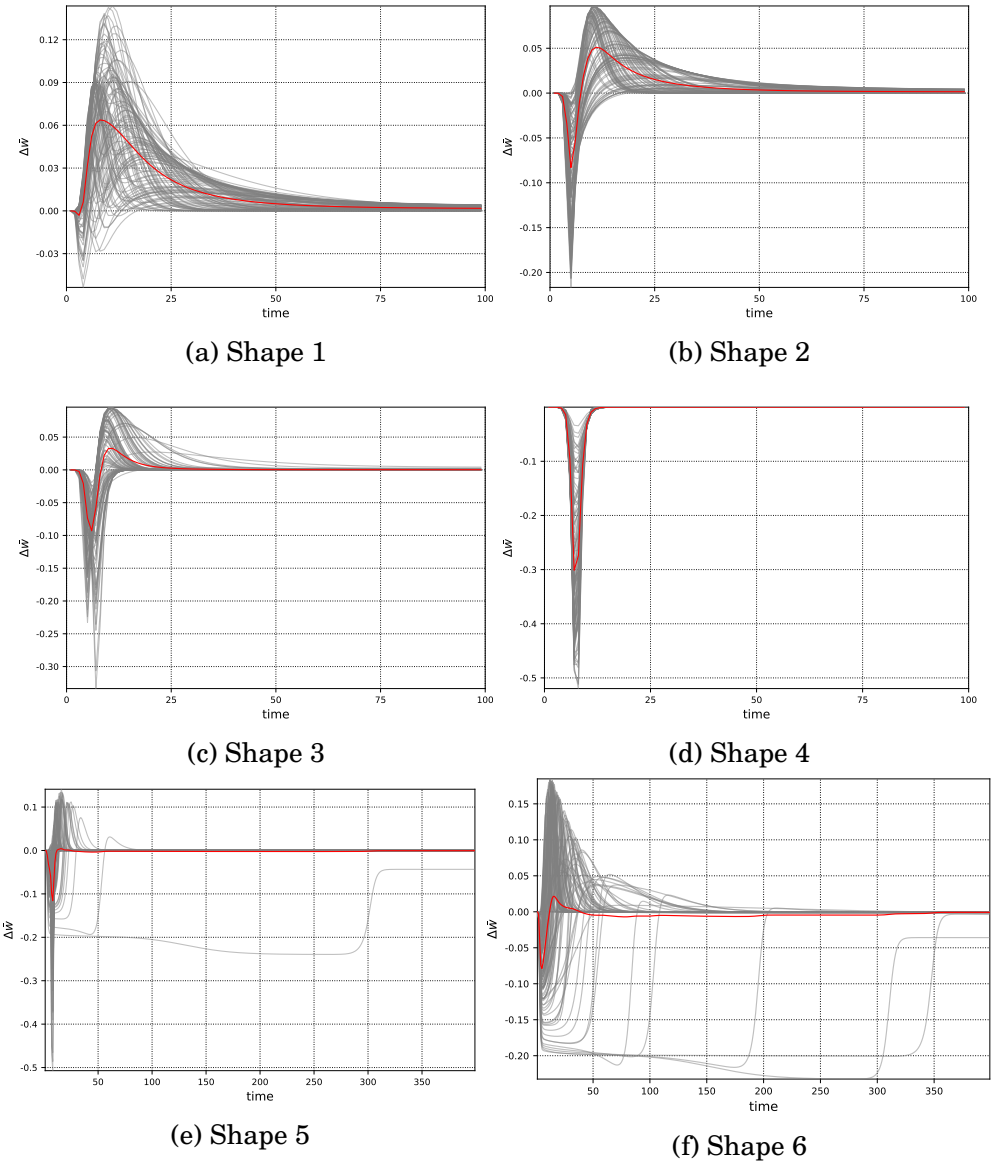
I later realized that by virtue of being an average quantity, $\Delta\bar{w}(t)$ lacks full information about the dynamics. Thus, in order to get more fine scaled results, I looked at the time variation of the difference of the frequencies of the triple mutants in either population i.e. $\Delta x_{111} = x_{111}^{\text{sex.}} - x_{111}^{\text{asex.}}$. As is evident from figure 6.10, shapes 1,2,3 and 4 landscapes have a fairly homogeneous response. However, considerable heterogeneity is seen for shape 5 and shape 6 landscapes. In both shapes 1 and 2, the recombining population performs better, even at long times. Despite having positive epistasis in the first half of the landscape, recombining populations on shape 2 landscapes still reach higher frequencies of x_{111}. In shape 3 landscapes there is a transitory advantage on average, but this declines over time. Similarly, in shape 4, there is a strong transitory disadvantage which declines over time. On average, recombination does not have any effect at long times in shape 5 and 6 landscapes, but in some specific landscapes, the triple mutant doesn't arise at all in the sexual population. Again, this is not surprising given that the disadvantage of recombination has already been noted in multi-peaked landscapes **[53]**.

Next, I considered modifier models to get conclusive results about the spread of recombination modifier alleles on PI landscapes of different shapes.

Modifier Model for PI landscapes

The following is a brief description of the modifier model:

1. It tracks the evolution of 4 locus sequences, wherein the fourth locus



(a) Shape 1

(b) Shape 2

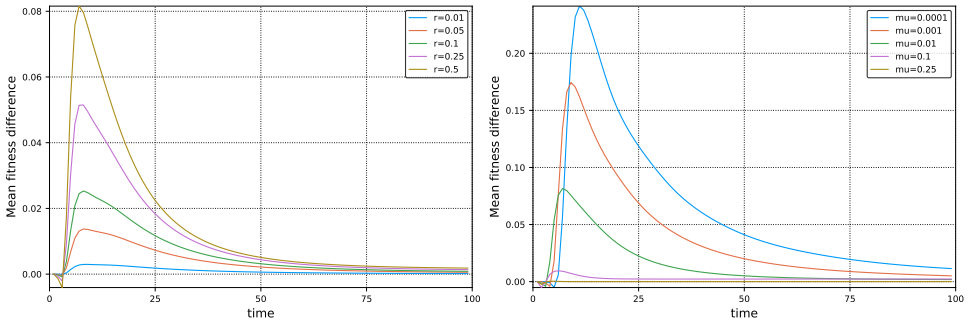
(c) Shape 3

(d) Shape 4

(e) Shape 5

(f) Shape 6

Figure 6.8: Variation of $\Delta\bar{w}(t)$ with time for all the 6 possible shapes of PI landscapes. $\mu = 0.01, r = 0.5$



(a) Variation with r.

(b) Variation with μ.

Figure 6.9: The mean of $\Delta\bar{w}(t)$ for shape 1 landscapes for different values of r and μ.

determines the recombination rate but doesn't contribute to the overall fitness of the genotype.

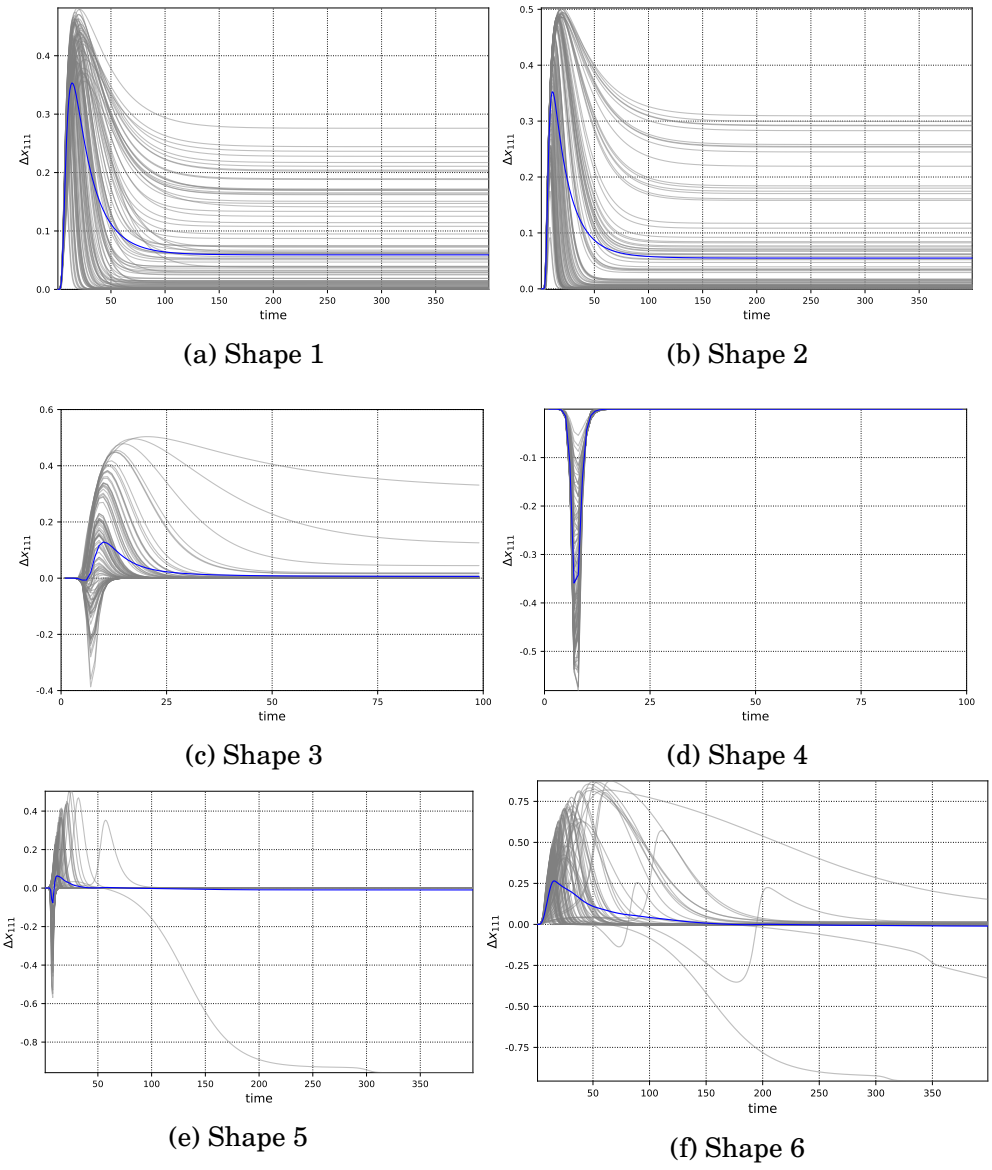
2. The fitness landscape of the first three loci is permutation invariant. An example of such a fitness landscape is shown in figure 6.11.

3. The initial population comprises only of sequences with the recombination modifier turned off (i.e. $\sigma_4 = 0 \Rightarrow$ an asexual population) and they are in mutation-selection equilibrium (with μ=0.1).

4. This equilibrium composition is then invaded in some fraction (here 0.5) by sequences with the recombination modifier turned on (i.e. $\sigma_4 = 1 \Rightarrow$ a sexual population). The role of the modifier locus is to modify the recombination rate (r) of two recombining sequences. If both recombining sequences have their modifier locus turned on, r=$2d_r$, where d_r is the modification rate. If only one of them has it on, r=d_r. Finally, if neither of them has it on, r=0.

5. The population then undergoes cycles of selection and recombination, until the stationary state is reached. In keeping with models found in literature, no mutations were considered post the invasion. This makes sense because the quenching of the effect of recombination by mutations is undesirable.

6. The aim is to study the dynamics of the allele frequency of the modifier locus (i.e. $\nu_4 = \sum_{\sigma:\sigma_4=1} x_\sigma$) and the frequency of the fittest genotypes (i.e. x_{1111} and x_{1110}) for all the 6 possible shapes.



(a) Shape 1

(b) Shape 2

(c) Shape 3

(d) Shape 4

(e) Shape 5

(f) Shape 6

Figure 6.10: Variation of triple mutant frequency difference with time for all the 6 possible shapes of PI landscapes.

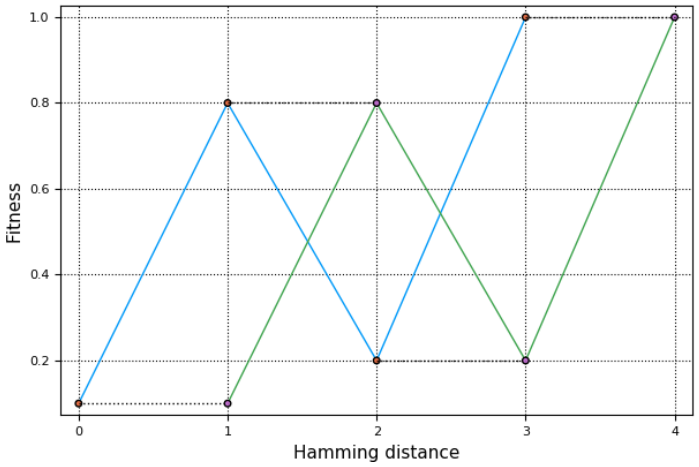
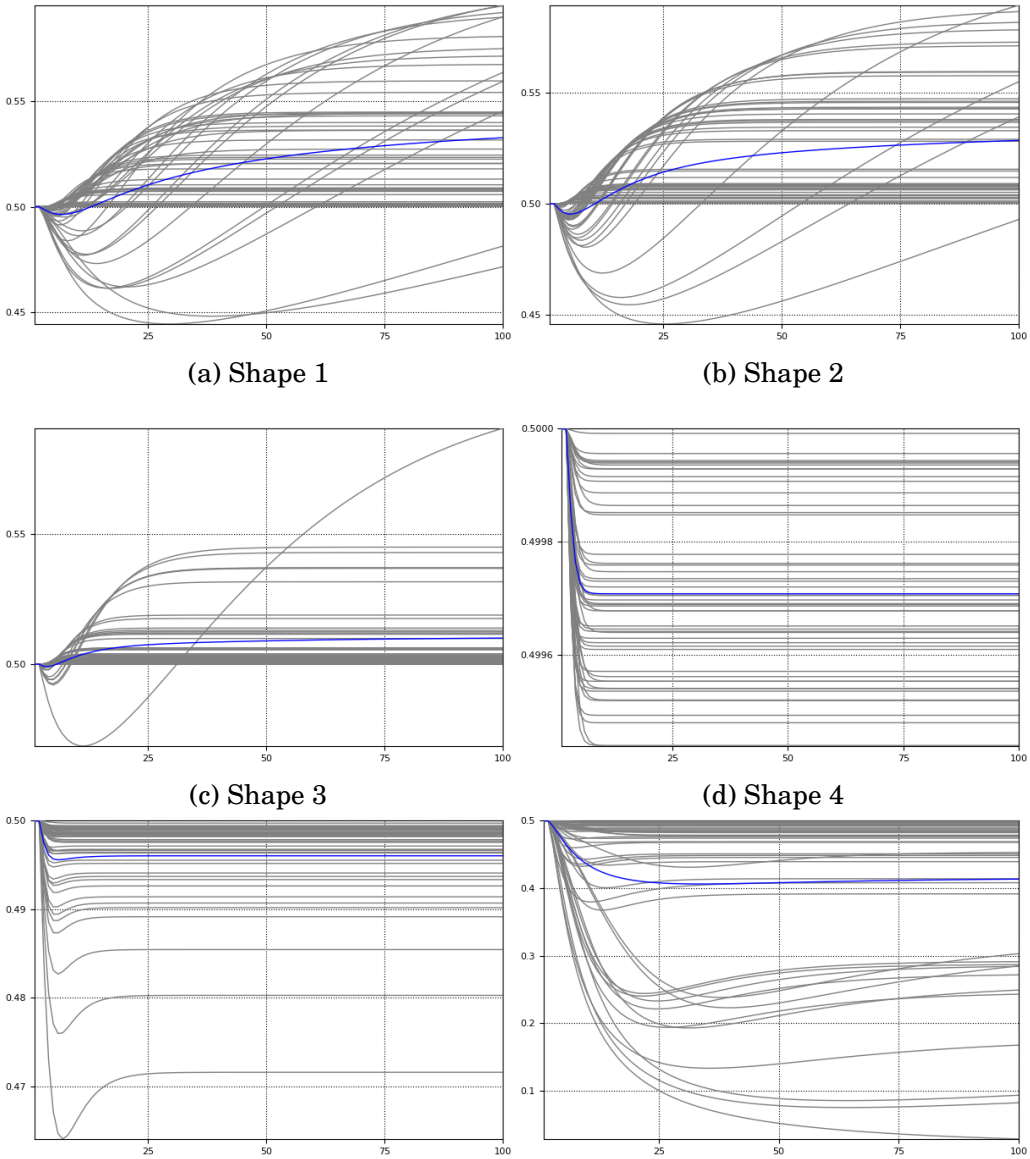


Figure 6.11: An example of a fitness landscape in the modifier model. The three locus sub-landscapes are of shape 6. The blue one is for sequences with $\sigma_4 = 0$, while the green one is for sequences with $\sigma_4 = 1$.

I studied the evolution of the modifier allele frequency, ν_4 for each of the shapes. The results are shown in figure 6.12. I started with equal fractions of asexual and sexual genotypes, meaning that $\nu_4(t = 0) =: a = 0.5$. Both the sexual and asexual populations were individually in mutation-selection equilibrium. Judging by the observed correlation between D and epistasis at equilibrium (figure 6.4), the signs of the different tests of D at the beginning of the dynamics are very likely to be equal to the sign of the corresponding epistasis tests. With the knowledge of the sign of D, one can already start guessing the outcome of the modifier dynamics for each shape. Since I'm ignoring mutations in the dynamics, the stationary state strongly depends upon the initial conditions. The initial conditions are determined by a and for every a there is a unique $\nu_4(t \longrightarrow \infty) = \nu_{\text{eq}}$. However, $\Delta\nu_4 = \nu_4^{\text{eq}} - a$ remains conserved and is the relevant quantity to study. The sign of $\Delta\nu_4$ is an indicator of whether recombination is advantageous or not. It turns out that $\Delta\nu_4$ is positive for shapes 1, 2 and 3 and negative for shapes 4, 5 and 6.

In figure 6.13, I have plotted the means of $\nu_4(\bar{\nu}_4)$ for each shape. This figure gives a clear picture about which shapes are advantageous for recombination.

I also looked at landscapes close to the transition point between shapes to see if any stark transition in the response curves occurs as one moves from one shape to the other. Figure 6.14 summarises that study by showing the variation of ν_4^{eq} i.e. $\nu_4(T_{\text{eq}})$ as a function of the shape of the landscape. Evidently, the stationary state of the dynamics shows scant regard for the shape



(a) Shape 1

(b) Shape 2

(c) Shape 3

(d) Shape 4

(e) Shape 5

(f) Shape 6

Figure 6.12: Time variation of v_4 for several landscapes of each possible shape. $b_r = 0$, $d_r = 0.25$, $\mu = 0$. Mean is shown in blue.

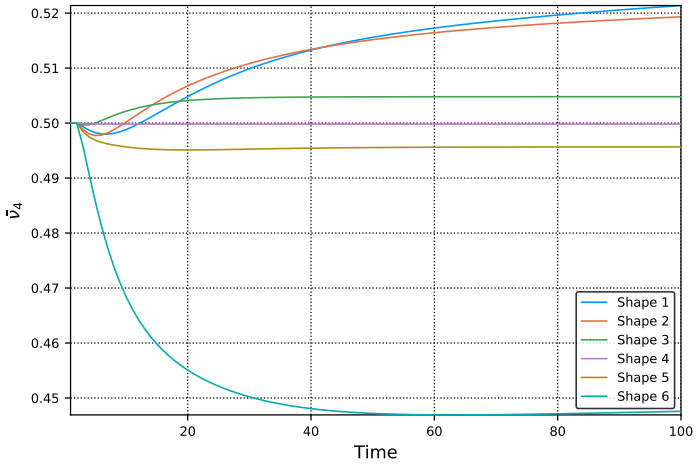


Figure 6.13: The inter-shape difference of the mean of the response curves

of the landscape on which the population evolves. Moreover, nothing exciting happens at the boundary between two different shapes. This leads us to conclude that the fate of the recombination modifier is not a characteristic of the shape of the fitness landscape.

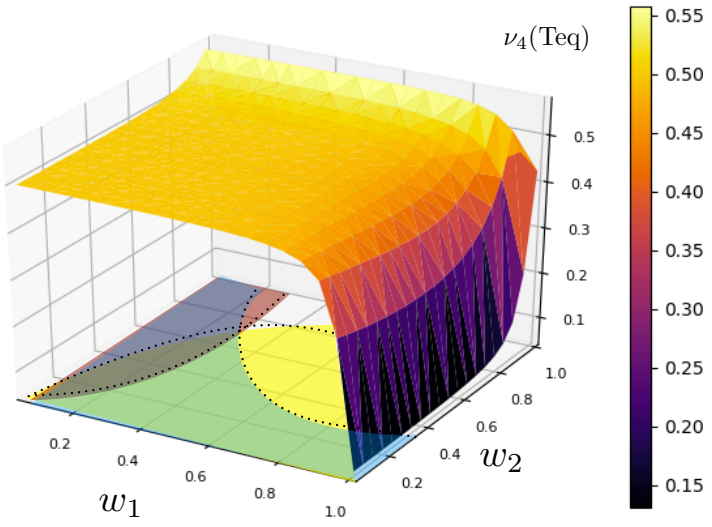


Figure 6.14: The variation of the stationary state of the modifier dynamics as a function of the shape of the fitness landscape. The dotted lines demarcate the different shapes.

Further, one must also note that the plots for shape 1 and 2 look nearly identical, although shape 1 has a larger variance, while on the other hand, shape 6 shows a very heterogeneous response. The extent to which the responses of shape 6 landscapes differ can be seen from figure 6.15. One may naively imagine the following two classes of PI landscapes to have different "shapes": $(f_0, f_1, f_2, f_3) = (0.1, 1-\alpha, 1-\alpha, 1)$ and $(f_0, f_1, f_2, f_3) = (0.1, 1-\alpha, 0.1+\alpha, 1)$, where $0 < \alpha <<< 1$. However, both landscapes have shape 6 and thus it is not surprising that their responses differ so much. This result implies that nothing concrete about the spread of ν_4 can be inferred from shape 6 landscapes.

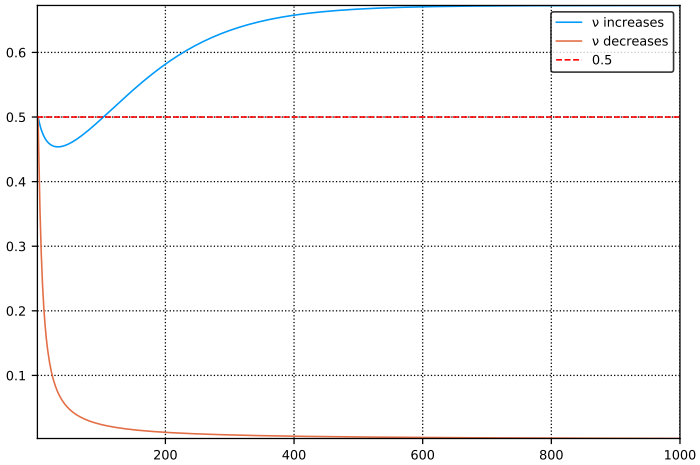


Figure 6.15: The heterogeneity in the response of shape 6 becomes evident from this plot. Plotted here for two shape 6 PI landscapes: (0.1, 0.99, 0.99, 1) in blue and (0.1, 0.99999, 0.11, 1) in orange.

Due to the initial diversity and the subsequent absence of mutations the populations end up reaching the global maxima, so at long times, $\Rightarrow x_{1111} + x_{1110} = 1$. The only interesting question is how the population is split between 1111 and 1110. One can qualitatively guess these results from the dynamics of ν_4 and also predict which shapes will be more heterogeneous.

Finally, I studied the variation of the stationary state (ν_4^{eq}) as a function of the recombination rate modifier (d_r) and the mutation probability (μ). The results for different shapes are shown in figures 6.16 and 6.17. Not surprisingly, $|\Delta\nu_4| = |\nu_4^{eq} - \nu_4(t=0)| = |\nu_4^{eq} - 0.5|$ (i.e. the magnitude of advantage or disadvantage of recombination) increases as d_r increases. The effect of switching on mutations is more interesting- $|\Delta\nu_4|$ first increases, then reaches a maximum at some intermediate value of μ and then begins to decrease to zero.

This makes sense because at sufficiently high mutation rates ($\mu > \mu_c$), the steady state of the population starts tending to a uniform distribution and thus $\nu_4^{eq} \to 0.5$. The initial increase can be interpreted as occurring due to rare mutations aiding the effect of recombination. Adding mutations also significantly increases the equilibration time.

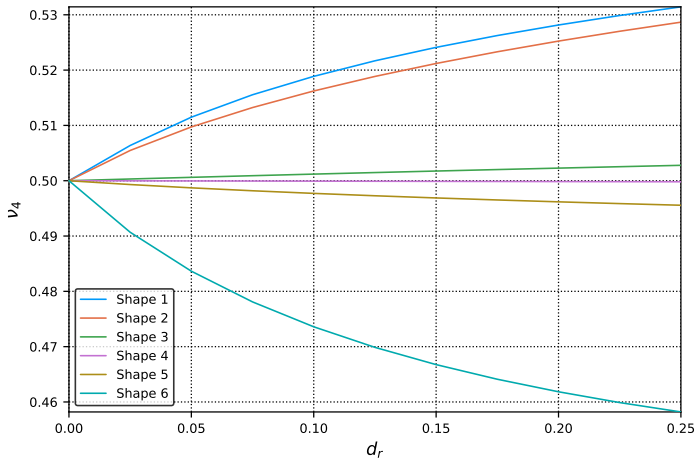


Figure 6.16: Dependence of the mean of the equilibrium modifier allele frequency on d_r, $\mu = 0$.

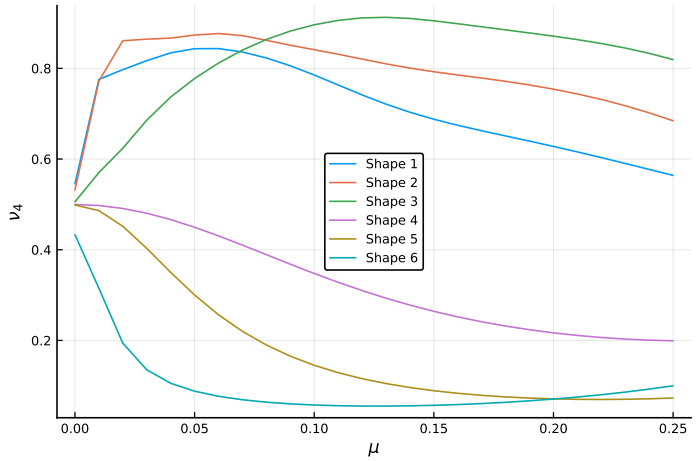


Figure 6.17: Dependence of the mean of the equilibrium modifier allele frequency on μ, $d_r = 0.25$. The sharp edges in the plot are because I changed μ in steps of 0.01.

Chapter 7

Final remarks

7.1 Conclusions

The idea of classifying fitness landscapes based on how they triangulate the Genotope is an intriguing one. Moreover, the connection that triangulations have to signs of circuits and Markov bases, allows them to be interpreted as summaries of epistatic interactions in the fitness landscape. At first glance, this interpretation makes the idea look promising and worth exploring.

However, the first thing to note is that the concept of shapes is only useful for 3 locus landscapes because they are almost trivial for 2 locus landscapes and unmanageably numerous for 4 locus landscapes and beyond. The 74 shapes of 3 locus landscapes can be classified into 6 types, but these types were initially difficult to understand. They became clearer on comparison with other more intuitive measures such as the number of peaks, sign epistasis motifs and the strength of higher order epistasis. The comparisons enabled us to learn that type 1 landscapes are on average most rugged, with maximum number of reciprocal sign epistasis motifs and maximum strength of higher order epistasis. On the other hand, type 6 landscapes have most correlated fitness values, causing them to be more likely to be single peaked.

After having gained more intuition about what the shapes mean, the next step was to assess the theory's utility. Broadly speaking, the theory can be used in two ways:

1. **Inferring epistatic interactions from theoretical and empirical fitness landscapes:** This is done by using the basis of the interaction space. This could either be the circuits or the Markov basis and they are particularly useful for multi-locus landscapes (even L>3) because they indicate which epistasis tests to consider. For combinatorially complete

fitness landscapes, the circuits are similar to Walsch coefficients, but they are more fine scaled, in the sense that the Walsch coefficients are averages of circuits. Circuits and Markov bases are also more useful because they can also be generated for landscapes with incomplete fitness values, which is usually the case for empirical fitness landscapes. The application of this theory to the three four locus β- lactamase landscapes enabled a more exhaustive study of the distributions of epistatic effects and the diminishing returns hypothesis. While the former revealed a peculiar, heavy tailed distribution for the landscape of large effect mutations, the latter seemed to hold very well for the landscape of synonymous mutations.

2. **Classifying landscapes into shapes, with the underlying assumption that same shapes share common properties:** As mentioned before, the classification is useful only for 2 and 3 locus landscapes, however shapes do provide some deeper insights.

   From the study of mutation-selection dynamics, we found that for two locus landscapes, the shape corresponding to negative epistasis has a larger equilibrium mean fitness ($\bar{w}_{\text{eq}}$) but also a longer time to equilibration (T_{eq}) than the shape corresponding to positive epistasis. We additionally found bounds on the $\bar{w}_{\text{eq}}$ and T_{eq} for fixed values of epistasis. Further, we found a sharpness in T_{eq} at exactly the transition point between the two shapes. For 3 locus landscapes, for $\mu < 0.1$, we found $\bar{w}_{\text{eq}}$ to increase on average with the type of the landscape, although the intra type variance was quite significant. On the other hand, the distribution of T_{eq} per type looks very different for small (~ 0.01) and large (~ 0.1) μ and also shows considerable variance. Here, for $\mu = 0.1$, we found T_{eq} to decrease with the type on average. These two results concord with the results on the average number of peaks in each type. The main message from the 3 locus mutation-selection study was that the pattern in which the fitness landscape decomposes the Genotope into tetrahedra does not have a very strong influence on the population dynamics on that landscape. This led us to add an additional constraint on the landscapes, namely permutation invariance.

   The analysis of the advantage of recombination for 3 locus PI landscapes led to a generalisation of Eshel and Feldman's two locus results– the three locus PI landscapes behaved like two independent two locus landscapes concatenated together, in the sense that sexual populations had a higher $\bar{w}(t)$ in the negatively epistatic parts of the landscape and vice

versa. This resulted in $\Delta x^{\text{eq}}_{111}$ being positive in shapes 1, 2 and 3 and negative in shapes 4, 5 and 6. This effect also reflected in the spread of the modifier allele in shapes 1, 2 and 3 and it being flushed out of the population in shapes 4, 5 and 6. The fact that different shapes showed characteristically different dynamics could be counted as a success of the shape theory, but at the same time one must note that the stationary state of the dynamics didn't show any strong dependence on the shape of the landscape. Added to that, shapes 1 and 2 had a very similar response, while shape 6 exhibited a very heterogeneous response. This hints that the shape is probably not the best way to classify landscapes. Finally, turning on mutations had a rather interesting effect – small μ enhanced the (dis)advantage of recombination, while large μ quenched it by trying to have equal proportions of sexual and asexual individuals.

To sum up, the shape theory definitely has relevance in the study of empirical fitness landscapes but when it comes to studying population dynamics, its use is confined to only 2 and 3 locus landscapes. Moreover, additional constraints are needed to actually infer something about the dynamics because the shape imposes too weak a constraint on the fitness landscape.

7.2 Future directions

There are a number of directions in which one can proceed from here.

On the experimental side, it will be interesting to firstly find out the biological reason behind the outlier epistatic interaction motif that was found for the landscape of large effect mutations. Additionally, one can test the predictions made by the shape of the landscape about the compositions of the fittest populations.

On the theoretical side, it remains to be seen if the different population dynamics models can be solved either exactly or approximately, in order to explain the simulation results. Perhaps, the most interesting will be to see if the small perturbation in the fitness landscape that lead to a drastically different response can be explained by looking at the stability of the fixed points. Further, one ought to explore how the results on the advantage of recombination change for finite populations, where there is more competition due to limited capacity. One can also study the advantage of recombination by comparing the time to adaptation for sexual and asexual populations.

Other potential applications of the shape theory can be in studying evolution by tracking the allele frequencies (by looking at paths in the Genotope)

and the disequilibrium measures or in studying the evolution of epistasis, by looking at the evolution of shapes as random adaptive walks on secondary polytopes. One can also look at shapes of empirical landscapes (preferably 3 locus landscapes), to see which are the most commonly occurring shapes in nature and to check if that can be explained by an evolutionary model of shapes.

Bibliography

[1] The nobel prize. https://www.nobelprize.org/prizes/chemistry/2018/summary/.

[2] M.V. Volkenstein. *Physical Approaches to Biological Evolution*. Springer-Verlag, 1994.

[3] Mutations. https://genetics.thetech.org/about-genetics/mutations-and-disease.

[4] H. Allen Orr. Fitness and its role in evolutionary genetics. *Nature reviews. Genetics*, 10:531–539, 2009.

[5] Hamming distance. http://www.oxfordmathcenter.com/drupal7/node/525.

[6] J. Arjan G.M. de Visser and Joachim Krug. Empirical fitness landscapes and the predictability of evolution. *Nature Reviews Genetics volume 15*, 15:480–490, June 2014.

[7] P. F. Stadler. Fitness landscapes. In *Lecture Notes in Physics, Biological Evolution and Statistical Physics*. Springer, Berlin, Heidelberg, 2002.

[8] S. Wright. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings of the Sixth International Congress of Genetics*, 1:356–366, 1932.

[9] S. A. Kauffman and S. Levin. Towards a general theory of adaptive walks on rugged landscapes. *Journal of Theoretical Biology*, 128:11–45, September 1987.

[10] S. A. Kauffman and E. D. Weinberger. The nk model of rugged fitness landscapes and its application to the maturation of the immune response. *Journal of Theoretical Biology*, 141:211–245, November 1989.

[11] Daniel M Weinreich, Yinghong Lan, C Scott Wylie, and Robert B. Heckendorn. Should evolutionary geneticists worry about higher-order epistasis? *Current opinion in genetics and development.*, 23:700–7, December 2013.

[12] P. C. Phillips. Epistasis: the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet. 9*, 9:855–67, November 2008.

[13] J. B. Wolf, E. D. Brodie, and M. J. Wade. *Epistasis and the evolutionary process*. Oxford University Press, 2000.

[14] Frank J. Poelwijk, Sorin Tănase-Nicola, Daniel J. Kiviet, and Sander J. Tans. Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes. *Journal of Theoretical Biology*, 272:141–144, 2011.

[15] Dmitry A. Kondrashov and Fyodor A. Kondrashov. Topological features of rugged fitness landscapes in sequence space. *Trends in Genetics*, 31:24–33, January 2015.

[16] EA Boyle, YI Li, and JK Pritchard. An expanded view of complex traits: From polygenic to omnigenic. *Cell*, 169:1177–1186, June 2017.

[17] J. Arjan G. M. de Visser, Tim F. Cooper, and Santiago F. Elena. The causes of epistasis. *Proceedings of the Royal Society B*, 278:3617–24, 2011.

[18] R. L. Malmberg. The evolution of epistasis and the advantage of recombination in populations of bacteriophage t4. *Genetics*, 86:607–621, 1977.

[19] Fyodor A. Kondrashov and Alexey S. Kondrashov. Multidimensional epistasis and the disadvantage of sex. *PNAS*, 98:12089–12092, 2001.

[20] Júlia Domingo, Guillaume Diss, and Ben Lehner. Pairwise and higher-order genetic interactions during the evolution of a trna. *Nature*, 558:117–121, 2018.

[21] Zachary R. Sailer and Michael J. Harms. High-order epistasis shapes evolutionary trajectories. *PLoS Comput Biol*, 13:e1005541, May 2017.

[22] Kristina Crona and Mengming Luo. Higher order epistasis and fitness peaks. *arXiv:1708.02063 [q-bio.QM]*, 2017.

[23] Frank J. Poelwijk, Vinod Krishna, and Rama Ranganathan. The context-dependence of mutations: A linkage of formalisms. *PLoS Comput Biol*, 12(6):e1004771, June 2016.

[24] Kristina Crona, Alex Gavryushkin, Devin Greene, and Niko Beerenwinkel. Inferring genetic interactions from comparative fitness data. *eLife*, 6:e28629, December 2017.

[25] Luca Ferretti, Benjamin Schmiegelt, Daniel Weinreich, Atsushi Yamauchi, Yutaka Kobayashi, Fumio Tajima, and Guillaume . Achaz. Measuring epistasis in fitness landscapes: The correlation of fitness effects of mutations. *Journal of Theoretical Biology*, 396:132–143, May 2016.

[26] Niko Beerenwinkel, Lior Pachter, and Bernd Strumfels. Epistasis and shapes of fitness landscapes. *Statistica Sinica, .*, 17(4):1317–1342, October 2007.

[27] Rekha R. Thomas. *Lectures in geometric combinatorics.*, volume 33 of *Student mathematical library, IAS/Park City mathematical subseries*. American Mathematical Society., 2006.

[28] Jesus De Loera, Joerg Rambau, and Francisco Santos. *Triangulations: Structures for Algorithms and Applications*. Springer-Verlag Berlin Heidelberg, 2010.

[29] Sergei Bespamyatnikh. Enumerating triangulations of convex polytopes. *Discrete Mathematics and Theoretical Computer Science Proceedings*, AA:111–122, 2001.

[30] Linear programming. http://demonstrations.wolfram.com/TheFundamentalTheoremOfLinearProgramming/.

[31] Peter Huggins, Bernd Sturmfels, Josephine Yu, and Debbie Yuster. The hyperdeterminant and triangulations of the 4-cube. *Mathematics of computation*, 77(263):1653–1679, July 2008.

[32] Jeanne Pellerin and Jean-Francois Remacle. Enumerating combinatorial triangulations of the hexahedron. *arXiv:1801.01288v2 [cs.CG]*, 2018.

[33] Roger D. Kouyos, Sarah P. Otto, and Sebastian Bonhoeffer. Effect of varying epistasis on the evolution of recombination. *Genetics*, 173:589–597, 2006.

[34] Peter Huggins, Lior Pachter, and Bernd Sturmfels. Towards the human genotope. *arXiv:q-bio/0611032v2 [q-bio.PE]*, 2006.

[35] Kristina Crona. Polytopes, graphs and fitness landscapes. In H. Richter and A. Engelbrecht, editors, *Recent Advances in the Theory and Application of Fitness Landscapes.*, volume 6. Springer, Berlin, Heidelberg, 2014.

[36] Niko Beerenwinkel, Lior Pachter, Bernd Sturmfels, Santiago F. Elena, and Richard E Lenski. Analysis of epistatic interactions and fitness landscapes using a new geometric approach. *BMC Evolutionary Biology*, 7:60, 2007.

[37] Sijmen Schoustra, Sungmin Hwang, Joachim Krug, and J. Arjan G. M. de Visser. Diminishing-returns epistasis among random beneficial mutations in a multicellular fungus. *Proc. R. Soc. B*, 283(1837), August 2016.

[38] Mark P. Zwart, Martijn F. Schenk, Sungmin Hwang, Bertha Koopmanschap, Niek Lange, Lion van de Pol, Tran Thi Thuy Nga, Ivan G. Szendro, Joachim Krug, and J. Arjan G. M. de Visser. Unraveling the causes of adaptive benefits of synonymous mutations in tem-1 β-lactamase. *Heredity*, 121:406–421, July 2018.

[39] Martijn F. Schenk, Merijn L.M. Szendro, Ivan G. Salverda, Joachim Krug, and J. Arjan G.M de Visser. Patterns of epistasis between beneficial mutations in an antibiotic resistance gene. *Molecular Biology and Evolution*, 30:1779–87, August 2013.

[40] Kavita Jain and Joachim Krug. Adaptation in simple and complex fitness landscapes. In *Structural Approaches to Sequence Evolution*. Springer, Berlin, Heidelberg, 2007.

[41] Eric W. Weisstein. Perron-frobenius theorem. http://mathworld.wolfram.com/Perron-FrobeniusTheorem.html.

[42] Hiroshi Furutani. Application of eigen's evolution model to infinite population genetic algorithms with selection and mutation. *Complex Systems*, 10:345–366, 1996.

[43] Alexander S. Bratus, Artem S. Novozhilov, and Yuri S. Semenov. Rigorous mathematical analysis of the quasispecies model: From manfred eigen to the recent developments. *arXiv:1712.03855v1 [q-bio.PE]*, 2017.

[44] Ilan Eshel and Marcus W. Feldman. On the evolutionary effect of recombination. *Theoretical Population Biology*, 1:88–100, May 1970.

[45] Non sister chromatids. https://teaching.ncl.ac.uk/bms/wiki/index.php/Non-sister%5Ctextunderscore%20chromatids.

[46] Daniel L. Hartl and Andrew G. Clark. *Principles of Population Genetics*. Oxford University Press, 2006.

[47] J. Arjan G. M. de Visser and Santiago F. Elena. The evolution of sex: empirical insights into the roles of epistasis and drift. *Nature Reviews Genetics*, 8:139–149, 2007.

[48] Sarah P. Otto and Thomas Lenormand. Resolving the paradox of sex and recombination. *Nature Reviews Genetics*, 3:252–261, 2002.

[49] Sarah P. Otto and Nick H. Barton. Selection for recombination in small populations. *Evolution*, 55:1921–1931, 2001.

[50] J.F. Crow and M. Kimura. *An introduction in Population Genetics Theory*. Harper and Row, New York., 1970.

[51] Su-Chan Park and Joachim Krug. Bistability in two-locus models with selection, mutation, and recombination. *J. Math. Biol.*, 62:763–788, 2011.

[52] Daniel M Weinreich, Suzanne Sindi, and Richard A . Watson. Finding the boundary between evolutionary basins of attraction, and implications for wright's fitness landscape analogy. *Journal of Statistical Mechanics: Theory and Experiment*, 2013, 2013.

[53] Stefan Nowak, Johannes Neidhart, Ivan G. Szendro, and Joachim Krug. Multidimensional epistasis and the transitory advantage of sex. *PLoS Computational Biology*, 10, 2014.