Processes Determining Genetic Variability:

Mutations in Sequence Space and Hitchhiking

D i s s e r t a t i o n

zur Erlangung des akademischen Grades

doctor rerum naturalium (Dr. rer. nat.),

vorgelegt dem Rat der biologisch–pharmazeutischen Fakultät

der Friedrich–Schiller–Universität Jena

von Thomas Wiehe

geboren am 16. Oktober 1961 in Coburg.

Gutachter:

- Prof. Dr. Peter Schuster (Friedrich–Schiller Universität, Jena)
- Prof. Dr. Wolfgang Stephan (University of Maryland at College Park, USA)
- Priv. Doz. Dr. habil. Gottfried Jetschke (Friedrich–Schiller Universität, Jena)

Tag des Rigorosums: 20. Dezember 1994

Tag der öffentlichen Verteidigung: 11. Januar 1995

Acknowledgements

I thank my advisors Peter Schuster and Wolfgang Stephan for providing an exciting research environment. Without them this thesis would not have been possible at all. I thank all my colleagues in Jena and Maryland for countless discussions, helps and hints. I thank Ellen Baake. A considerable part of this work has been shaped by collaboration with her. I thank Volker Wünsche. He opened the door to science for me. I thank my dear friends Alessandra, Clemente and Mariella. They have also helped me through some less than easy times. I thank Nana who had the stars in her eyes. I thank my parents. Their support has never been in question and their patience was infinite.

Contents

Introduction

Already in the early days of theoretical genetics the topic of genetic variability led to an intense dispute about its role in evolution. Is variability advantageous or detrimental to a species? For decades scientists argued in favor of one (Dobzhansky, 1955) or the other (Muller, 1950) view mainly for theoretical and plausibility reasons. The arguments of both sides rested upon the assumption that some form of selection was the principal evolutionary evolutionary force. But even when in the sixties experimental data started to become available (Lewontin & Hubby, 1966), the problem could not be resolved clearly.

Nowadays it seems obvious, however, that the question about advantage or disadvantage of variability – at least in this naive way – had not been posed in the right way. Different aspects have to be taken into account.

Of course, variability is the raw material without which evolution could not proceed. It is just as clear that too much of variation destroys exactly what is to be conserved: The basic information encoding the blue print of an organism, which has to be passed on from generation to generation. Replication of DNA has to be sufficiently accurate to ensure the 'informational' and thus the actual survival of a species. Eigen (1971) presented a model which incorporates these ideas in the context of viral and of prebiotic evolution of biological macromolecules.

The presence of genetic variability is a fact for any natural and also artificial species (cf. the experiments of *in vitro* evolution by Spiegelman *et al.* (1965)).

Genetic variants are generated by mutations. They may be somatic or affect the germ line. Here, we are exclusively concerned with the latter ones. Furthermore, they may be due to replication inaccuracy or to some external impact. This difference, as will become clear, does essentially not affect our results.

Likewise recombination has an enhancing effect on variability. Mutation works gradually. Mutants accumulate step by step (Muller, 1964; Haigh, 1978), if back-mutations are ignored. In contrary to this, recombination may create large genetic distances between wild-type and recombinant on the spot. It may also serve as an effective means to re-create a lost wild type; this is virtually impossible to accomplish by mutation alone.

Whether selection enhances or reduces variability depends on the type of selection. Balancing

selection is sometimes believed to play an important role for sustaining polymorphism at certain gene sites; most prominent are perhaps studies concerning the polymorphism of the *MHC* complex (Takahata *et al.*, 1992). On the other hand, directional selection leads to an elimination of one or the other allele from the population. These types of selection are discussed within the framework of two–allele models. More sophisticated selection models are constructed on sequence space. Here, fitness values are ascribed to particular nucleotide sequences. This fitness assignment may originate from energy properties of the secondary structure associated with the (RNA-) nucleotide sequence. Such folding landscapes have been introduced by Fontana & Schuster (1987).

Looking more closely at data on molecular variation one observes that, within a genome, there are quite different levels of variation present.

A genome may thus not be viewed as a homogeneous entity, but it appears that – possibly due to the function of different genes – variability is in some regions enhanced above an average level, whereas in others it is reduced. What are possible mechanisms for such an irregular pattern? Clearly, some combination of the above forces will be responsible. Advantageous mutants – on their way to fixation in a population – may wipe out variability in their neighborhood: Reduction of variation by hitchhiking.

Our starting point will be the classical mutation selection equation in the context of haploid reproduction which – by a suitable interpretation of the involved parameters – serves as a model of prebiotic evolution. An important feature is the phenomenon of the error threshold which has been studied extensively during the last decades (Swetina & Schuster, 1982; Nowak & Schuster, 1989). This feature carries over to the biotic context of evolution of natural organisms. Threshold phenomena are also observed if the mutation selection equation is put into the framework of replication of diploid individuals.

Striking as the evidence for presence of error thresholds appears for a certain type of fitness assignments, we ask in the second chapter whether this threshold phenomenon can also be observed for more general fitness landscapes.

In the third chapter, we extend the original model to include the evolutionarily important feature of recombination. This procedure leads into the realm of so called 'two locus' systems, which are of course a standard object since the early theoretical studies in population genetics. In the first three chapters we study model systems from the theoretical point of view. All occuring parameters are phenomenological. The link to application is put up in the fourth chapter. Here, we introduce a model for genetic hitchhiking – again a two locus system.

In the fifth chapter we apply this model to data on genetic variability of natural populations and aim on estimating evolutionary parameters from these data.

In chapter six we give a short outlook into limitations and problems of the discussed models and a possible strategy for improvement.

Finally, in the appendix we elaborate on some more technical mathematical derivations.

Repeatedly used symbols are collected on page 8.

Part of this work has been carried out with Wolfgang Stephan at the University of Maryland, the other part with Peter Schuster at the Institute for Molecular Biotechnology in Jena. This thesis cannot – and does not intend to – deny its heterogeneous character. Chapters four and five are due to collaboration with W. Stephan. The models there have their origin in experimental investigations of molecular mechanisms of nucleotide variation in the genome of natural organisms.

On the other hand, chapters one through three originated from the theoretical question to which extent properties of the quasi-species model, the genuine context of which had been prebiotic and viral evolution, carry over into models of diploid reproduction.

Chapter three tries to show the importance of recombination for evolution of diploid organisms. This feature has not been treated in the original models of replicating systems (as adopted in chapters one and two). On the other hand, it is an essential ingredient when it comes to describe variability (chapters four and five). When dealing with large genome sizes – as in higher organisms – it becomes indispensable to take the possibility of recombination into account. In other words, chapter three intends to serve as a stepping stone from prebiotic to biotic models.

We try to indicate how concepts from part one may serve to improve models to attack the problem of part two: What are the mechanisms generating genetic variability?

Symbols

$\mathcal{A}$	First gene locus in two locus model	(p. 48)
$\mathcal{B}$	Second gene locus in two locus model	(p. 48)
$\mathcal{H}$	$:= \bar{H}(\infty)$, equilibrium nucleotide heterozygosity	(p. 82)
$\mathcal{L}$, $\mathcal{L}^*$	Differential operators	(p. 71)
A_i, A, a, B, b	nucleotide strings or alleles	(p. 11)
D	Linkage disequilibrium	(p. 49)
E, $E(p)$, $E(\mu)$	Average Hamming distance of population from master sequence with respect to the stationary distribution (therefore dependent on the mutation term p (μ))	(p. 19)
$E(Y_0)$	Expectation of Y_0 with respect to $\phi_p(y)$	(p. 26)
H, $H(t)$	Heterozygosity at locus $\mathcal{A}$ $\left(:= 2y_A(t)(1-y_A(t)) \right)$	(p. 69)
$\bar{H}(t)$	Weighted average of heterozygosity	(p. 69)
H_{neut}	$:= H(0) = \bar{H}(0)$, neutral heterozygosity (without influence of selection at neighboring site)	(p. 78)
$I_{R^*}(\alpha)$	An abbreviation for the integral in Eq.(5.1)	(p. 82)
L	Superimposed single peaked and neutral fitness landscape	(p. 37)
L_H	Multiplicative fitness landscape	(p. 32)
L_{SP}	Single peaked fitness landscape	(p. 31)
L_δ^Δ	Fitness landscapes taking arbitrary fitness values between δ and Δ	(p. 35)
$L_{\hat{\nu}}$	Fitness landscapes modeling truncation selection: $v_i = 0$ for all $i > \hat{\nu}$	(p. 35)
M	Mutation matrix with entries m_{ij}	(p. 12, 33)
$\hat{M}$	Mutation matrix with entries $\hat{m}_{ij}$ (Poisson approximation of M)	(p. 33)
N	Population size	(p. 24)
R	Recombination rate	(p. 49)

R_{max}	Critical recombination rate (single peaked landscape)	(p. 52)
R_{min}	Critical recombination rate (directional selection)	(p. 64)
$\bar{R}$	Maximal recombinational distance within which hitchhiking is effective	(p. 78,82)
R^*	$:= 2N\bar{R}$	(p. 82)
T_*, T^*	Net recovery times for master A after a substitution at locus $\mathcal{B}$ took place	(p. 56)
V, $V(p)$, $V(\mu)$	Variance of Hamming distance	(p. 19)
W	Symmetric fitness matrix with entries w_{ij}	(p. 12)
$(Wx)_i$, w_i	$\sum_{j=1}^{n} w_{ij}x_j$ marginal fitness of gamete A_i	
(x, Wx), $\bar{w}$	$= \sum_{j=1}^{n} w_j x_j$ mean fitness of diploid population	(p. 12)
Y_0	The random variable 'Relative frequency of master allele'	(p. 25)
Y_1	The random variable 'Relative frequency of error tail'	(p. 25)
$Z_{A\|B}$, $Z_{A\|b}$	The random variable 'Relative frequency of A conditioned on its occurrence with B (b)'	(p. 74)
$a(y)$	Drift coefficient of diffusion process	(p. 25)
$b(y)$	Diffusion coefficient of diffusion process	(p. 25)
c_1, c_2	Coefficients of linear regression model	(p. 86)
$d(A_i, A_j)$	Hamming distance between two strings A_i and A_j (=number of nucleotides in which they differ)	(p. 11)
f	$= 1 - \epsilon^{\frac{R}{\ln 2}}$	(p. 62)
h	Dominance parameter	(p. 15)
$h_{red}(\rho m)$	Reduction in heterozygosity due to a single substitution event at distance between m and $m + dm$ from locus $\mathcal{A}$	(p. 78)
m_{ij}	Mutation probability. With probability m_{ij} string A_j mutates into one of form A_i	(p. 12)
n	$= \kappa^\nu$ number of possible alleles of sequences of length ν	(p. 12)
p	Single digit (=nucleotide) mutation probability	(p. 13)
p_{max}	Critical mutation probability	(p. 21)
p^{max}	Critical mutation probability (2nd definition)	(p. 43)
q	$= 1 - p$. Single digit accuracy	(p. 11)
s	Selective differential between master and non-master alleles	(p. 15)
s_1, s_2	Selective differentials at loci $\mathcal{A}$ and $\mathcal{B}$, respectively	(p. 53)
t	Variable denoting time	(p. 12)
v_i	Fitness of (haploid) genotype A_i	(p. 14)

(v, x), $\bar{v}$	$= \sum_{j=1}^{n} v_j x_j$ mean fitness of haploid population	(p. 14)
w_{ij}	Fitness of (diploid) genotypes $A_i A_j$	(p. 12)
$\bar{w}_{rel}$	Relative excess of diploid mean-fitness ('advantage of sex')	(p. 22)
x_i	Allelic frequency (of allele A_i) in one locus model ($i \in \{1, ..., n\}$)	(p. 12)
	Gametic frequency in two locus two allele model ($i \in \{1, ..., 4\}$)	(p. 49)
y_i	Sum of relative frequencies of alleles carrying i mutations with respect to a previously specified master allele (='error class' or 'Hamming class' frequency)	(p. 16)
y_A, y_a, y_B, y_b	Relative frequencies of master (A, B) alleles and error tails (a, b) in two locus model	(p. 50)
$z_{A\|B}$	Relative frequency of master A conditioned on its occurrence with master B	(p. 68)
$z_{A\|b}$	Relative frequency of master A conditioned on its occurrence with error tail b	(p. 68)
α	$= 2Ns$	(p. 26, 75)
θ	Rate of occurrence of selected substitutions at locus $\mathcal{B}$	(p. 61)
$\hat{\theta}$	Critical rate of occurrence of selected substitutions at locus $\mathcal{B}$	(p. 64)
κ	Length of the alphabet (=number of types of nucleotides)	(p. 12)
λ	Genome mutation rate. Related to μ by $\lambda = \mu\nu$	(p. 32)
μ	Single digit mutation rate	(p. 14)
μ_{max}	Critical mutation rate	(p. 21)
ν	Number of nucleotides in particular stretch of DNA or RNA	(p. 11)
$\nu_{\mathcal{A}}$	Number of nucleotides at $\mathcal{A}$–locus in two locus model	(p. 48)
$\nu_{\mathcal{B}}$	Number of nucleotides at $\mathcal{B}$–locus in two locus model	(p. 48)
π	Nucleotide diversity as obtained from experimental studies; here to identify with $\mathcal{H}$	(p. 82)
ρ	Per nucleotide recombination rate	(p. 78)
τ	Variable denoting time	(p. 12)
$\phi_p(y)$	Stationary density of diffusion process	(p. 26)
$\phi(.,.,t)$	Transition density of diffusion process in two locus model	(p. 74)
χ	Rate of occurrence of selected substitutions which drag the linked allele to fixation	(p. 78)
ψ	Average number of selected substitutions per nucleotide site per generation	(p. 78)
$\tilde{.}$	. to be replaced by w_{ij}, W, m_{ij}, M; yields then the analogous expressions for the decoupled equations	(p. 12)

1

Effect of Multiple Mutation

Mutation is an unavoidable side reaction of reproduction or the inevitable consequence of some external impact. Although the molecular structure of DNA had been known since 1953 (Watson & Crick, 1953) it received attention in population genetics only in the late sixties, when it was explicitly taken into account in models which included the action of mutation (Jukes & Cantor, 1969; Eigen, 1971). In the latter one, correct replication and mutation were considered as parallel biochemical reactions within the same general mechanism. In this framework, polynucleotide strings such as DNA or RNA are identified with points in sequence space. This is, in case of binary strings, the discrete hypercube of dimension ν, where the integer ν is the (fixed) length of the string considered. A natural metric on the hypercube is the Hamming metric $d(A_i, A_j)$, which counts the number of different positions in two aligned strings A_i and A_j. Mutation frequencies were introduced in a somewhat simplified way that was later on characterized as the uniform error rate model (for a review see Eigen *et al.* (1989)). They are fully determined by an accuracy parameter (q), the chain length (ν) of the polynucleotide, and the distance in sequence space between the correct copy and the mutant which is tantamount to the Hamming distance between the two sequences.

The dynamics of the frequency distribution of alleles in a population under the interplay of correct and erroneous replication as well as selection may be viewed as a process of *error propagation*, an outstanding feature of which is the loss of genetic information when a critical error rate is surpassed.

For diploid organisms, mutation–selection equations have been studied in the more general context of population genetics (i.e., without reference to sequence space) as 'Fisher's equation with mutation' in two different versions (Crow & Kimura, 1970; Hadeler, 1981). In this chapter we focus on these equations with the mutation terms adapted to sequence space. Depending on the properties of the selection matrices, which are determined by the fitness landscape, properties of the haploid sequence space models carry over, or new and unexpected behavior may emerge.

1.1 Model Equations

Mutation–selection equations were formulated in two different versions, depending on the assumptions on the mutation mechanism: If mutations originate as replication errors on the occasion of reproduction events, the appropriate ODE system for the composition of an (infinite) diploid population with overlapping generations in Hardy-Weinberg equilibrium (cf. Hofbauer & Sigmund (1988)) reads

$$\dot{x}_i = \sum_j (m_{ij}\, x_j (Wx)_j) - x_i (x, Wx)\ , \quad i = 1, \ldots, n\,. \tag{1.1}$$

Here, n is the number of alleles ($n = \kappa^\nu$, where e.g. $\kappa = 2$ in binary sequence space, $\kappa = 4$ in case of natural DNA or RNA sequences), and the x_i denote the relative frequencies of alleles A_i (with $\sum_i x_i = 1$). W is the symmetric $n \times n$ matrix of fitnesses (to be interpreted as reproduction rates) of (ordered) A_iA_j genotypes. $(Wx)_i := \sum_{j=1}^n w_{ij}x_j$ is the marginal fitness of allele A_i and $(x, Wx) := \sum_{j,k=1}^n w_{jk}x_jx_k$ denotes the average fitness of the population. It is assumed that all genotypes have identical death rates; so, due to normalization, these do not contribute to the equations. M is the mutation matrix with elements $m_{ij} := m_{i \leftarrow j}$ denoting the *probability* that an A_j allele is replicated as A_i, with $\sum_i m_{ij} = 1$.

Let us recall that the assumption of Hardy-Weinberg equilibrium is an approximation in the case of overlapping generations, even if random mating is assumed (Hadeler, 1974). Equation (1.1), called the 'coupled mutation selection equation' in what follows, was first studied by Hadeler (1981).

With the transformation of time: $d\tau := (x(t), Wx(t))dt$, $\tau = \int_0^t (x(\theta), Wx(\theta))d\theta$ may be interpreted as the time scale of generations, with 'generation' understood as the time span it takes the population to grow to e times its size (e is Euler's constant), where deaths are not counted. This yields

$$\frac{dx_i}{d\tau} = \frac{\sum_j m_{ij}\, x_j (Wx)_j}{(x, Wx)} - x_i\,. \tag{1.2}$$

A formally identical equation has been arrived at in passing from the difference equation for non-overlapping generations to an ODE (Ewens, 1979).

If one relaxes the assumption that mutation occurs at the time of replication but stipulates that it may occur at any time in the life cycle of a gamete, mutation and selection must be considered as independent events. One is then led to the following decoupled version of the selection mutation equation

$$\dot{x}_i = x_i\{(\tilde{W}x)_i - (x, \tilde{W}x)\} + (\tilde{M}x)_i \quad i = 1, \ldots, n\,, \tag{1.3}$$

which was formulated in (Crow & Kimura, 1970), and thoroughly investigated by Akin (1979). Here, $\tilde{m}_{ij}$ denotes the mutation *rate* from A_j to A_i for $i \neq j$, and $\tilde{m}_{ii} = -\sum_j \tilde{m}_{ji}$. The entries in the symmetric fitness matrix $\tilde{W}$ are now to be interpreted as Malthusian fitnesses; in particular,

the $\tilde{w}_{ij}$ need no longer be positive. However, the equation is invariant under the transformation $\tilde{w}_{ij} \to \tilde{w}_{ij} + c$ for an arbitrary constant c, if performed for all index pairs i, j simultaneously. Thus, it is no restriction to consider $\tilde{W}$ to be a positive matrix as well.

Hofbauer (1985) discusses the relationship between the coupled and decoupled versions. In particular, he shows that Eq. (1.3) is the limiting version of Eq. (1.1) in the limit of small selective differentials and small mutational terms. As far as error thresholds are concerned, mutational terms need, however, not be sufficiently small as to neglect the difference between the two models.

From the biological point of view, the relative contributions of the underlying mutation mechanisms are controversial. A mechanism described by Eq. (1.1) implies constancy of mutation rates per generation, whereas Eq. (1.3) is suited to describe systems with mutation rates which are constant per year. 'Real life' evolution does, of course, have contributions from both replication errors and independent processes like radiation damage that become manifest at replication time but occur with constant intensity in time. The – quite controversial – experimental evidence concerning the proportions of these contributions has been discussed for example by Kimura (1987). In this paper the cautious conclusion is drawn that a constant rate per year, with a certain bias from generation time, may be a realistic assumption. In contrary to this, Ohta (1993) reported recently significant generation time effects for several mammalian species. After all, the whole molecular clock analysis hinges on the assumption that mutations occur at rates close enough to constancy in time.

This issue still being unresolved clearly, we analyze both versions in a parallel fashion.

As mentioned above, we identify 'alleles' with nucleotide sequences of length ν. The mutation matrices M and $\tilde{M}$ are specified to be of sequence space type. For the coupled model this means

$$m_{ij} = \left(\frac{p}{\kappa - 1}\right)^{d(A_i,A_j)} (1-p)^{\nu - d(A_i,A_j)} , \tag{1.4}$$

where p is the mutation probability per nucleotide. We assume that mutations occur independently at any site and that, given a mutation at a site occurs, any other nucleotide is equally likely to be incorporated instead of the original one.

The mutation matrix for the decoupled equation, on the other hand, must exclude double mutations, i.e.,

$$\tilde{m}_{ij} = \begin{cases} \frac{\mu}{\kappa-1} & (d(A_i,A_j) = 1) , \\ -\nu\mu & (d(A_i,A_j) = 0) , \\ 0 & \text{otherwise} . \end{cases} \tag{1.5}$$

Here, μ is the single digit mutation *rate*.

In the sequel, we will concentrate on binary sequence space and assume $\kappa = 2$. This assumption does not severely restrict the validity of our investigations since any alphabet may be translated into a binary one.

Let us briefly comment on how to compare p and μ. In the case $w_{ij} \equiv v_o$ (no selection),

Eq.s (1.1) and (1.3) reduce to

$$\dot{x} = v_o(M - I)x \tag{1.6}$$

and

$$\dot{x} = \tilde{M}x\,, \tag{1.7}$$

respectively, where I denotes the identity matrix. $\tilde{M}$ and $M-I$ are symmetric, they commute and may thus be diagonalized simultaneously. From the explicit knowledge of the spectral properties of M as analyzed by Rumschitzky (1987), it follows that $v_o(M - I)$ and $\tilde{M}$ have corresponding eigenvalues $v_o\{(1-2p)^{\nu-\rho}-1\}$, and $2\mu(\rho-\nu)$, respectively, both with multiplicity $\binom{\nu}{\rho}$, $\rho = 0, \ldots, \nu$. So, obviously, Eq. (1.7) is the first order perturbation expansion in p of Eq. (1.6) when p is identified with μ/v_o.

In order to compare haploid and diploid models, it is essential to know for which fitness schemes the diploid equations reduce to their haploid counterparts. For Eq. (1.3), it is clear that, with the fitness of heterozygotes (*arithmetically*) intermediary between the corresponding homozygotes, i.e., $\tilde{w}_{ij} = \frac{1}{2}(\tilde{w}_{ii} + \tilde{w}_{jj})$, the dynamics reduces to that of a haploid population. For, in this case, the difference between marginal fitness of allele A_i and mean fitness is

$$(\tilde{W}x)_i - (x, \tilde{W}x) = \frac{1}{2}(\tilde{w}_{ii} - \sum_j \tilde{w}_{jj}\, x_j), \tag{1.8}$$

so

$$\dot{x}_i = x_i(\tilde{v}_i - (\tilde{v}, x)) + (\tilde{M}x)_i\,. \tag{1.9}$$

Here, $\tilde{v}_i := \frac{1}{2}\tilde{w}_{ii}$ may be interpreted as the fitnesses of single alleles which make up a haploid population and $(\tilde{v}, x)$ as the corresponding mean fitness $\sum_i \tilde{v}_i x_i$.

For the coupled equation (in the form (1.2)), on the other hand, the diploid dynamics reduces to the haploid case, if heterozygote fitness equals the *geometric* mean of the corresponding homozygote fitness values, i.e. $w_{ij} = \sqrt{w_{ii}w_{jj}} = v_i \cdot v_j$. In this case, we have

$$\frac{dx_i}{d\tau} = \frac{\sum_j m_{ij} v_j x_j}{(v, x)} - x_i\,. \tag{1.10}$$

1.2 Equilibria

It has been shown (Thompson & McBride, 1974; Jones *et al.*, 1976) that, via the transformation $z_i = x_i \exp(\int_0^t (v, x(\tau))d\tau)$, the second degree ODE's describing the coupled haploid replication dynamics can be transformed into a set of linear differential equations. Thus, the problem to solve the haploid system reduces to the spectral problem of the matrix $M \cdot (\mathrm{diag}(v))$. Furthermore, for $p > 0$ it follows from the positivity of $M \cdot (\mathrm{diag}(v))$ and the Perron-Frobenius theorem (see for example (Karlin & Taylor, 1975; Heuser, 1992)) that the ODE system has a unique, globally stable equilibrium in the interior of the $n-1$ dimensional unit simplex of relative frequencies.

The same results hold for the decoupled equations, as the above considerations go through for $\tilde{M} + (\mathrm{diag}(\tilde{v}))$ instead of $M \cdot (\mathrm{diag}(v))$.

A similar transformation to reduce the degree of the system is not known for the diploid case. Even worse, the number of equilibria and their stability properties depend crucially upon the entries of the fitness matrix W and $\tilde{W}$, respectively (detailed analyses may be found in e.g. (Karlin, 1980; Kingman, 1988)). For the decoupled equation, fitness matrices may be constructed s.t. oscillating behavior of the trajectories emerges (Akin, 1979). The resulting fitness landscapes have more than one maximum on the simplex in all cases considered. We will show shortly that even for fitness landscapes with unique maxima multiple equilibria may occur, which rules out the possibility of global convergence even for such situations.

1.3 Effects of Dominance on Error Thresholds

In haploid sequence space models, error thresholds were studied extensively on single-peaked fitness landscapes (Nowak & Schuster, 1989), as characterized by one 'fit' master allele with competitors of reduced but equal fitness (here we use the term 'master allele' synonymously with 'allele with highest fitness').

It is well known that in this case the (unique) stationary distribution of relative frequencies exhibits a sharp transition into a distribution which is approximately uniform as the nucleotide mutation probability surpasses a critical value p_{max}, which depends on both, the selective advantage of the master and the chain length ν. Such a transition has been identified as an *error threshold* e.g. by Swetina & Schuster (1982).

As a well-suited diploid analogue to the haploid single-peaked landscape, we propose the following model. All genotypes are classified according to whether they are

a. homozygous with both alleles carrying no mutations (fitness f_0),

b. heterozygous with one allele carrying an arbitrary number of mutations and the other carrying none (fitness f_1) or

c. homozygous with both alleles carrying mutations (fitness f_2).

Here, 'carrying mutations' and 'carrying no mutations' are to be understood with respect to the previously distinguished master allele. We express the f_i in terms of the selective advantage s and a dominance parameter h:

$$
\begin{aligned}
f_0 &= w_{11} = (1+s)^2\,, \\
f_1 &= w_{1i} = w_{i1} = (1+s)^{2h}\,,\ i \neq 1 \text{ and} \\
f_2 &= w_{ij} = 1\,,\ i,j \neq 1
\end{aligned}
$$

and for the decoupled equations

$$f_0 = \tilde{w}_{11} = 1 + 2s\,,$$
$$f_1 = \tilde{w}_{1i} = \tilde{w}_{i1} = (1 + 2hs)\,,\ i \neq 1 \text{ and}$$
$$f_2 = \tilde{w}_{ij} = 1\,,\ i,j \neq 1\,.$$

For $0 \leq h \leq 1$, this defines a single-peaked landscape in the above sense. In particular, for $h = \frac{1}{2}$ the fitness regimes reduce to the multiplicative and additive cases, respectively. It is worth mentioning that both fitness schemes coincide to first order in s.

Like its haploid counterpart, the diploid single-peaked landscape allows reduction of systems (1.1) and (1.3) to ODE systems of $\nu + 1$ equations for $y_0, \ldots, y_\nu$, where y_i denotes the relative frequency of alleles carrying i mutations with respect to the master allele (i.e. relative frequency of *Hamming class* i). In general, this reduction is possible if $d(A_i, A_1) = d(A_k, A_1)$ together with $d(A_j, A_1) = d(A_l, A_1)$ imply $w_{ij} = w_{kl}$ for all $i,j,k,l \in \{1, \ldots, n\}$. This *homogeneity condition* is clearly satisfied in case of our single peaked landscape. The mutation matrices have to be adjusted in the following way

$$m_{ij} = (1-p)^\nu \sum_{k=j-\nu+i}^{\min(i,j)} \binom{j}{k}\binom{\nu-j}{i-k}\left(\frac{p}{1-p}\right)^{i+j-2k} \qquad (i,j = 0,\ldots,\nu)\,, \tag{1.11}$$

and

$$\tilde{M} = \mu \begin{pmatrix} -\nu & 1 & & & 0 \\ \nu & -\nu & 2 & & \\ & \ddots & \ddots & \ddots & \\ & & 2 & -\nu & \nu \\ 0 & & & 1 & -\nu \end{pmatrix}. \tag{1.12}$$

Let us resort to the single-peaked example to discuss the equilibrium properties mentioned previously. Bürger (1983) has shown that, in cases where the selection vector field alone yields a globally stable stationary state, the decoupled mutation selection equation may have more than one stable equilibrium if the mutation matrix is non-symmetric. We will show now that counterexamples to global stability with single-peaked landscapes may also be found when mutation matrices are of sequence space type (and therefore symmetric in the formulation of (1.1) or (1.3)). Let us start with the slightly simpler decoupled equation.

With the choice $h = 0$, the equilibrium condition (in Hamming class notation) reduces to

$$(\tilde{M}y)_i = 2sy_0^2 \cdot \begin{cases} (y_0 - 1)\,, & i = 0 \\ y_i\,, & \text{otherwise} \end{cases} \tag{1.13}$$

Now consider the inhomogeneous system of linear equations arising from Eq. (1.13) by formally replacing $2sy_0^2$ by α:

$$\tilde{M}y = \alpha(y - e_1) \tag{1.14}$$

with parameter α, where $e_1 := (1, 0, \ldots, 0)^T$. As long as $\alpha > 0$, it has a unique solution $\bar{y}(\alpha)$ (since the eigenvalues of $\tilde{M}$ are ≤ 0) with components of the form

$$\bar{y}_i(\alpha) = \frac{P_i(\alpha)}{Q(\alpha)}\,,\ i = 0, \ldots, \nu\,, \tag{1.15}$$

where $P_i(\alpha), Q(\alpha)$ are polynomials in α of degree at most $\nu + 1$. Specializing to $\nu = 4$ and substituting $\alpha = 2sy_0^2$, we obtain equilibrium values of y_0 as roots of the following polynomial $P(y_0)$, defined as

$$\begin{aligned} P(y_0) &:= y_0 Q(2sy_0^2) - P_0(2sy_0^2) \\ &= -3 + 48y_0 - 32\sigma y_0^2 + 100\sigma y_0^3 - 40\sigma^2 y_0^4 \\ &\quad +70\sigma^2 y_0^5 - 16\sigma^3 y_0^6 + 20\sigma^3 y_0^7 \\ &\quad -2\sigma^4 y_0^8 + 2\sigma^4 y_0^9\,, \end{aligned}$$

where $\sigma := s/\mu$. We have thus reformulated the problem of finding solutions of Eq. (1.14) to finding roots of a 9th degree polynomial. The remaining entries of $(y_0, \ldots, y_\nu)^T$ are then uniquely determined by Eq. (1.15). We have $P(0) < 0$ and $P(1) > 0$, independently of σ. Choosing $\sigma = 8$, $P(y_0)$ is easily seen to have at least three roots in $[0, 1]$. We have thus constructed a single-peaked situation with (at least) three equilibria.

For the coupled equation, the situation is hardly more complicated. The equilibrium condition now reads (with $\alpha := (2 + s)sy_0^2$)

$$(My)_i = (1 + \alpha)y_i - \alpha\, m_{i0}\,,\ i = 0, \ldots, \nu\,. \tag{1.16}$$

Again, situations with multiple equilibria are obtained – e.g. $\nu = 4$, $p = 1/20$, and $s = 9/20$. It can be shown that (at least) two of them are stable.

In order to study diploid error threshold behavior, we numerically follow branches of equilibria as the mutation rate is increased. We start with $p = 0$ ($\mu = 0$, respectively) and a homogeneous population consisting of master strings only. The curves in Figure 1.1 represent stationary frequencies of Hamming classes. A transition into the (approximate) uniform distribution for $p > p_{max}$ is evident.

In analogy to the haploid case we call such a transition 'error threshold'. However, in the diploid situation the location of the transition depends on the dominance parameter h.

For $h = \frac{1}{2}$, haploid and diploid thresholds coincide. On the other hand, recessivity ($h < \frac{1}{2}$) or dominance ($h > \frac{1}{2}$) of the master cause shifts of the error threshold with respect to the haploid calibration in opposite directions; see Figure 1.2, and Table 1 for numerical values (for the latter, we used equilibrium master frequency $< 1\%$ as the threshold criterion). In the case of (total) dominance ($h = 1$), the critical mutation rate is roughly doubled with respect to the haploid case, and the transition becomes smoother; recessivity, on the other hand, shifts the threshold

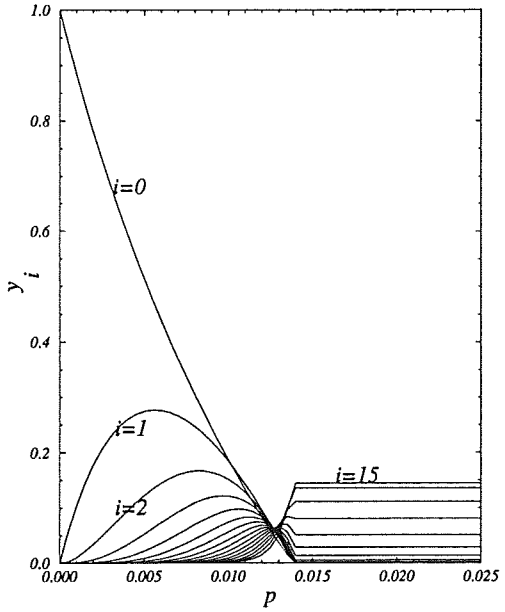

Figure 1.1: Stationary distribution of Hamming classes $i = 0$ to $i = 15$ in dependence of nucleotide mutation probability. Diploid case. Parameters: $s = 1/2$; $h = 1/2$; $\nu = 30$. Plotted are equilibrium frequencies y_i of Hamming classes.

to lower values and sharpens it. A master which is dominant in heterozygotes can tolerate much higher mutation rates before it is lost from the population (at equilibrium) than a master which is intermediary or recessive in heterozygotes. We will come back to this point later.

The possible existence of multiple equilibria implies a dependency of the dynamics on initial conditions. We illustrate this point briefly and single out two scenarios.

a) An advantageous allele (master) preexists in an initially homogeneous population and accumulates point mutations as time proceeds until mutation selection balance is reached, i.e. $y_0(0) = 1$.

b) The population is initially uniformly distributed in sequence space, i.e. $y_i(0) = \binom{\nu}{i}/2^\nu$.

Think that in case b) the advantageous allele A_0 has been introduced into the population by a spontaneous mutation just recently, such that it is present at time $t = 0$ only in low frequency. Figure 1.4 illustrates how the equilibrium distribution depends on the two different initial conditions. Obviously, the error threshold is shifted to a lower value in case b). In general, it can be shown that bistability (i.e. two biologically meaningful, stable equilibria) may occur only if $h < 1/3$. For such h–values the interval of mutation rates, within which bistability is possible, extends over $[\mu_1, \mu_2] = [\frac{2sh}{\nu}, \frac{(1-h)^2 s}{2\nu(1-2h)}]$. Thus, the dominance parameter h induces a bifurcation of equilibria. A more thorough discussion of this point is in preparation (Baake & Wiehe, 1995).

Approximate values for the location of the threshold can be derived in a way similar to that employed by Eigen (1971). We consider a caricature of the original model by neglecting back-flow

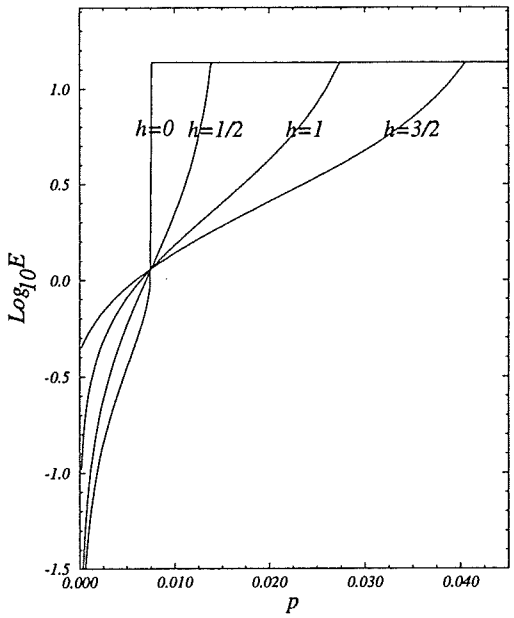

Figure 1.2: Average Hamming distance from master $E = E(p)$ at stationarity. Parameters: $s = 1/2$; $\nu = 30$; $h = 0, 1/2, 1, 3/2$. Thresholds are located at p values, where average hamming distance enters constancy. E saturates at $\nu/2$.

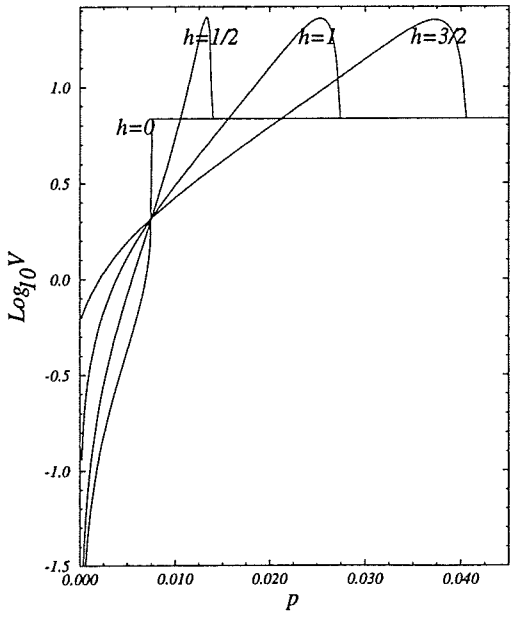

Figure 1.3: Variance of Hamming distance from master ($V = V(p)$) at stationarity. Parameters: $s = 1/2$; $\nu = 30$; $h = 0, 1/2, 1, 3/2$.

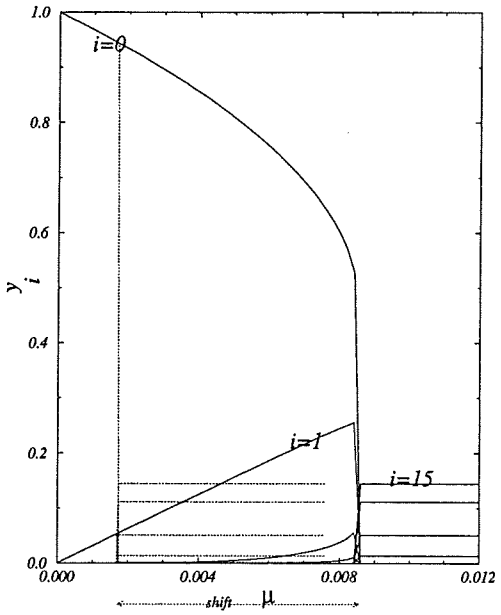

Figure 1.4: Depending on initial conditions, different equilibria may be attained. A threshold shift (indicated by arrow "shift") to a lower value is induced by passing from scenario a) to scenario b) (see text). Above graphics shows the situation for the decoupled equations. Qualitatively the same holds for the coupled system. Plotted are equilibrium frequency distributions of scenario a) (solid lines) and scenario b) (dotted lines) as μ varies. Parameters: $\nu = 30$; $s = 1/2$; $h = 1/20$. The lower threshold ($\mu = 1.67 \cdot 10^{-3}$) is predicted by analytical formula (1.21).

mutation and concentrate on the differential equation for the master frequency x_1. At stationarity, we are left with

$$0 = x_1(Wx)_1 m_{11} - x_1(x, Wx)\,. \tag{1.17}$$

Its nontrivial solution $\bar{x}_1 \neq 0$ is characterized by

$$m_{11} = \frac{(x, Wx)}{(Wx)_1}\,, \tag{1.18}$$

where $m_{11} = (1-p)^\nu$, $(Wx)_1 = x_1(1+s)^2+(1-x_1)(1+s)^{2h}$, and $(x, Wx) = 1+4hsx_1+2sx_1^2(1-2h)$. We find that $\bar{x}_1 > 0$ as long as

$$p < p_{max} := 1 - \left(\frac{1}{(1+s)^{2h}}\right)^{\frac{1}{\nu}}\,, \tag{1.19}$$

and $\lim_{p\to p_{max}} \bar{x}_1 = 0$.

The corresponding quantity for the haploid case is

$$p_{max} = 1 - \left(\frac{1}{1+s}\right)^{\frac{1}{\nu}}\,. \tag{1.20}$$

Similarly, for the decoupled system (1.3) one obtains

$$\mu_{max} = \frac{2sh}{\nu}\,. \tag{1.21}$$

In case of underdominance ($h < 0$), the above expressions (Eq.(1.19), (1.21)) have to be taken as identical 0. If one has bistability, these threshold formulae correspond to the lower threshold (scenario b)).

Crude as these approximations certainly are, they demonstrate that, in the coupled case, the ratio of homo- and heterozygote non-master fitness is a relevant quantity, whereas it is their difference in the decoupled case. With fitness parameters to be interpreted as reproduction rates in both cases, the above relations are consistent with the interpretation of μ as a mutation *rate* (and μ_{max} a difference of rates) and p as a mutation *probability* (and p_{max} a ratio of rates, hence dimension-less). Also, it is consistent with the invariance properties of the equilibria: Equilibria of Eq. (1.1) are invariant with respect to $w_{ij} \to w_{ij} \cdot c$, whereas those of (1.3) are invariant with respect to $w_{ij} \to w_{ij} + c$ for all i, j.

We give a list of threshold values for a choice of parameters in Table 1.1. Dominance does not only affect the location of the error threshold, but it also tunes the variability of a population. If population quantities, like *average Hamming distance from master* or *variance of Hamming distance from master* are compared for different values of h and plotted as functions of p (similar results hold for the decoupled equation), one observes an inflection at $p = \widehat{p}$ and three characteristic parameter regions (see Figures 1.2 and 1.3): A region of very exact replication $p \in I_0 = [0, \widehat{p}]$, an intermediary one with $p \in I_1 = [\widehat{p}, p_{max}]$ and the region beyond the error threshold I_2 with $p > p_{max}$. For small p, variability at equilibrium (measured in terms of expectation and variance

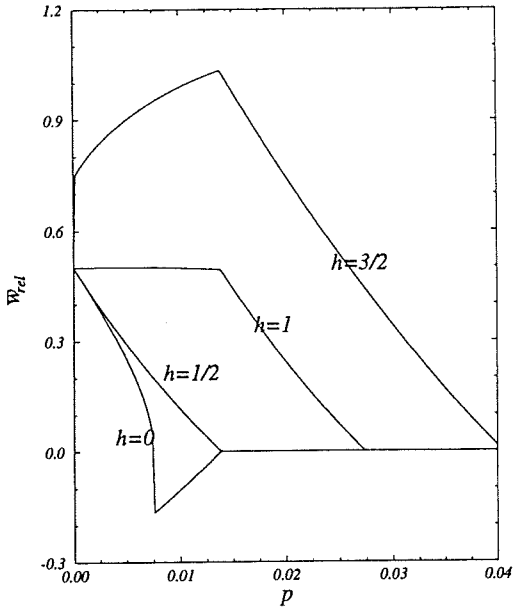

Figure 1.5: Advantage of diploidy $\bar{w}_{rel}$ for $h = 0, 1/2, 1, 3/2$ in dependence of p. Parameters: $\nu = 30,\ s = 1/2$.

of Hamming distance) is higher for dominant master alleles and lower for recessive ones. For intermediary p ($p \in I_2$) the situation is reversed. Here, dominance leads to a relatively lower variability and a higher concentration of the population around the master allele. Whereas for $p > \widehat{p}$ dominance 'shields' the master from a quick loss from the population, for values $p < \widehat{p}$ recessivity is needed to keep the master allele at highest possible concentrations. We calculated the inflection point $\widehat{p}$, where the effects of dominance and recessivity are reversed to be approximately

$$\widehat{p} = 1 - \left(\frac{\sqrt{1+2s}-1}{s}\right)^{\frac{1}{\nu}} . \tag{1.22}$$

Expanding in a Taylor series and retaining terms of order up to $O(s^2)$ one derives

$$\widehat{p} \approx \frac{s}{2\nu}(1+\frac{s}{4}) . \tag{1.23}$$

Comparing *mean fitness* for different values of h yields a different picture, however. Following Chao (1988), we relate diploid and haploid mean fitnesses by the quantity $\bar{w}_{rel}$ (called 'advantage of sex' in Chao (1988)), which is

$$\bar{w}_{rel} = \bar{w}_{rel}(h) := \frac{(x, W(h)x) - (v, x)}{(v, x)} . \tag{1.24}$$

In the entire domain of p values, we find $\bar{w}_{rel}(h_1) < \bar{w}_{rel}(h_2)$, if $h_1 < h_2$. Furthermore, if $p < p_{max}(h)$ one has $\bar{w}_{rel}(h) \geq 0$ for all h, indicating an advantage of diploidy as long as mutation probabilities are sufficiently small (see Figure 1.5).

1.4 Time Dependent Behavior

For the haploid dynamics the eigenvalue spectrum of the matrix $M \cdot (\mathrm{diag}(v))$, plotted in dependence on p, shows a series of 'avoided crossings' as pointed out by Nowak & Schuster (1989). In particular, the avoided crossing of the first and second largest eigenvalues coincides with the error threshold, indicating a singularity of convergence times of the dynamics in this region, since the ratio of eigenvalues λ_1 and λ_2 approaches unity. In the diploid replication regime, a similar explosion of convergence times is observed. Due to the non–linearity of order three of systems (1.1), (1.2) and (1.3) an eigenvalue characterization of convergence velocity is, however, not readily available in the diploid case.

We present results of numerical integrations of system (1.2) (for comparability with the haploid case) and analytical approximations. For the latter we again neglect back-flow terms. We discuss the coupled equation in the following; very similar results are obtained when the decoupled equation is treated along the same lines. For the master in Eq. (1.2), one has the following rational first order autonomous ODE

$$\frac{dx_1}{d\tau} = \frac{x_1(Wx)_1 m_{11}}{(x, Wx)} - x_1 \,. \tag{1.25}$$

We concentrate on the initial value $x_1(0) = 1$. To derive convergence times, it suffices to integrate the reciprocal of the right hand side of (1.25) with respect to x_1. This yields

$$\tau = \int_1^{\bar{x}_1+\epsilon} \frac{(\xi, W\xi)}{\xi(Wx)_1 m_{11} - \xi(\xi, W\xi)} d\xi \,. \tag{1.26}$$

Thus, τ gives the time it takes for the master to reach its equilibrium value $\bar{x}_1 + \epsilon$ starting from $x_1(0) = 1$.

Although the integral in Eq. (1.26) can be evaluated for arbitrary h, for simplicity we present here only the result for $h = 1/2$ explicitly. In this case we derive

$$\tau = \frac{m_{11}(1+s)\log(\frac{1+s(\bar{x}_1+\epsilon)-m_{11}(1+s)}{1+s-m_{11}(1+s)}) - \log(\bar{x}_1+\epsilon)}{1-m_{11}(1+s)} \,. \tag{1.27}$$

The right hand side of (1.27) has a singularity at the threshold value $p = p_{max}$. This fact again emphasizes the important role of the error threshold. For the mutation force being that strong, evolution comes practically to a standstill, no matter how the composition of the population in detail is. This is true also for other choices of h.

In Figure 1.6, we show convergence times for different initial conditions as obtained by numerical integration of the entire system (1.2). To this end, we first integrated the ODE system over a long period of time ($t = 10^5$ proved adequate for our choice of parameters in Figure 1.6) to determine equilibria $\bar{x}$. In a second run, we measured the time the system needed to approach $\bar{x}$ s.t. $|x_i - \bar{x}_i| \leq \epsilon$ for all i, with $\epsilon = 10^{-5}$. Important appears the fact that convergence times have a minimum for some value p_{min} with the property $0 < p_{min} < p_{max}$, if 'worst case' initial conditions are assumed, i.e. $x_n(0) = 1$ (in this case the master can only be generated by back-mutations).

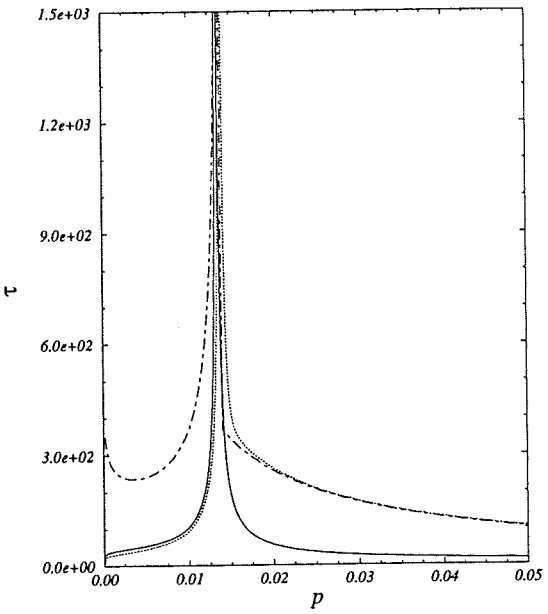

Figure 1.6: Time τ to reach equilibrium. Solid line: Analytical expression (1.27). Initial condition $x_1(0) = 1$. Dotted: Numerical integration. Initial condition $x_1(0) = 1$. Dashed: Numerical integration. Initial condition $x_n(0) = 1$ ('worst case condition'). Numerical and analytical curves agree very well for $p < p_{max} = 0.013$. For p beyond the threshold back-flow mutations become important, which have been neglected in the analytical treatment. This explains the discrepancy of numerical and analytical solutions in this region. Parameters: $\nu = 30$, $s = 1/2$, $h = 1/2$, $\epsilon = 10^{-5}$.

A new advantageous allele, out-competing all other alleles and at a maximum distance from the current – homogeneous – population composition is established most quickly for definitely positive mutation rates around p_{min}. In a rapidly changing environment fast convergence to new equilibria is an essential requirement for quick adaptation.

Intuitively, one could expect that a dominant master allele is faster at establishing itself in the population than is an intermediary or even recessive one. In contrast to this supposition we find that convergence times – for the caricature as well as for the entire system – do in fact depend only very weakly on any dominance properties, except near the poles.

1.5 Finite Populations

So far, the effect of random genetic drift has been totally neglected. Dealing with reproduction of diploid organisms in a general setting, the effects of finite population size (N in the haploid case, $2N$ in the diploid case) on the composition of the gene pool can in general not be ignored, especially when considering sequence space. Here, even at moderate chain lengths, the number of possible individuals by far exceeds any realistic population size. We investigate the influence upon the error threshold due to finite population size by modeling stochasticity as classical multinomial

Wright–Fisher sampling (cf. Ewens (1979), chpt. 3). The according transition to a continuous state space leads to a time homogeneous diffusion process. As previously, we perform numerical simulations of the entire system, along with an analytic treatment of a simplified model. In both cases, we are interested in stationary distributions. For these to exist, none of the states may be absorbing. Instead of neglecting back-flow altogether, we now resort to the simplification introduced in (Nowak & Schuster, 1989). We collect all alleles except the master into a single class, called 'error tail', and assume all different alleles to be equidistributed within it. We thus obtain reduced mutation matrices

$$M = \begin{pmatrix} (1-p)^\nu & \frac{1-(1-p)^\nu}{2^\nu-1} \\ 1-(1-p)^\nu & 1-\frac{1-(1-p)^\nu}{2^\nu-1} \end{pmatrix}, \tag{1.28}$$

and

$$\tilde{M} = \begin{pmatrix} -\nu\mu & \frac{\nu\mu}{2^\nu-1} \\ \nu\mu & -\frac{\nu\mu}{2^\nu-1} \end{pmatrix}. \tag{1.29}$$

We emphasize that the former reduction from originally n to $\nu+1$ equations leaves the dynamics of the Hamming classes unaffected as long as the homogeneity condition is met. On the other hand, this second reduction to only *one* equation (as suggested by the single peaked landscape) is an approximation.

Mutation terms m_{ij} now have a third meaning (cf. pp. 13 and 16). However, it is clear from the context which one is intended.

We consider a one-dimensional diffusion process, representing the relative frequency Y_0 of the master allele in the population. The frequency of the error tail is simply $Y_1 = 1 - Y_0$. Y_0 and Y_1 are random variables on the unit interval $[0,1]$. Infinitesimal drift and diffusion coefficients $a(y)$ and $b(y)$ are those for a two-allele model in population genetics and their derivation is standard (see e.g. (Ewens, 1979)). For our landscape $w_{00} = (1+s)^2$, $w_{01} = w_{10} = (1+s)^{2h}$, $w_{11} = 1$ (as well as for its approximation $w_{00} = 1+2s$, $w_{01} = w_{10} = 1+2hs$, $w_{11} = 1$) they read

$$a(y) = 2sy(1-y)(y+h(1-2y)) + (1-y)m_{01} - ym_{10} \tag{1.30}$$

as drift and

$$b(y) = \frac{1}{2N}y(1-y) \tag{1.31}$$

as diffusion coefficient. Specializing to $h = \frac{1}{2}$ and substituting N for $2N$ yields drift and diffusion for the haploid case with fitness scheme $v_0 = 1+s$ and $v_1 = 1$. Constraints on the order of magnitude of the parameters s, m_{01} and m_{10} as discussed by Ethier & Kurtz (1986), (Chpt. 10.1) are met, if $w_{ij} = 1 + O(N^{-1})$ and $m_{ij} = O(N^{-1})$. These are technical requirements in order to establish the diffusion process with coefficients $a(y)$ and $b(y)$ as the limit in distribution of the underlying discrete Markov chain. For practical purposes, however, it turns out that these constraints can be relaxed. The diffusion approximation will be satisfactory if $(1+2s)^{-1} \approx 1-2s$.

The dominance parameter h is not affected by this type of constraint. Error thresholds are determined by analysis of the stationary distribution of the random variable Y_0. The stationary density function $\phi_p(y)$ is obtained as equilibrium solution of the Kolmogorov forward equation (the index p emphasizes its dependence on nucleotide mutation probability p via m_{01} and m_{10}):

$$\phi_p(y) = cy^{4Nm_{01}-1}(1-y)^{4Nm_{10}-1}e^{4Nsy(y+2h(1-y))}, \tag{1.32}$$

where c is the normalization constant (see e.g. Ewens (1979); his parameter s has to be replaced by $2s$ due to the different selection regime used here).

ϕ_p may be shown to perform a transition into a function monotonously decreasing in p as a certain value p_0 is surpassed. For a similar problem (Nowak & Schuster, 1989), this transition was taken as a criterion to diagnose an error threshold in finite populations. In our case, however, we find this criterion to severely misestimate thresholds for wide ranges of parameters. We therefore propose to study directly the expectation of the random variable Y_0,

$$E(Y_0) = \int_0^1 y\phi_p(y)dy\,. \tag{1.33}$$

In analogy to the deterministic case, we locate the threshold where – at equilibrium – the master is 'lost' from the population. In the case of finite populations 'loss of the master' is naturally identified with $E(Y_0) < \frac{1}{2N}$. For consistency with the deterministic calculations, however, we prefer to consider the master to be lost as soon as its expected relative frequency is smaller than 1%.

For certain values of h, $E(Y_0)$ may be calculated explicitly.

$$h = 0: \qquad E(Y_0) = \gamma\,\frac{{}_2F_2((\frac{1+\beta_2}{2}, 1+\frac{\beta_2}{2}), (\frac{1+\beta_1+\beta_2}{2}, \frac{2+\beta_1+\beta_2}{2}), 2\alpha)}{{}_2F_2((\frac{1+\beta_2}{2}, \frac{\beta_2}{2}), (\frac{1+\beta_1+\beta_2}{2}, \frac{\beta_1+\beta_2}{2}), 2\alpha)} \tag{1.34}$$

$$h = \frac{1}{2}: \qquad E(Y_0) = \gamma\,\frac{{}_1F_1(1+\beta_2, 1+\beta_1+\beta_2, 2\alpha)}{{}_1F_1(\beta_2, \beta_1+\beta_2, 2\alpha)} \tag{1.35}$$

$$h = 1: \qquad E(Y_0) = \gamma\,\frac{{}_2F_2((\frac{1+\beta_1}{2}, \frac{\beta_1}{2}), (\frac{1+\beta_1+\beta_2}{2}, \frac{2+\beta_1+\beta_2}{2}), -2\alpha)}{{}_2F_2((\frac{1+\beta_1}{2}, \frac{\beta_1}{2}), (\frac{1+\beta_1+\beta_2}{2}, \frac{\beta_1+\beta_2}{2}), -2\alpha)} \tag{1.36}$$

where $\gamma = \frac{\beta_2}{\beta_1+\beta_2}$, $\alpha = 2Ns$, $\beta_1 = 4N(1-(1-p)^\nu)$ and $\beta_2 = 4N\frac{(1-(1-p)^\nu)}{2^\nu-1}$. Furthermore, ${}_1F_1$ denotes the confluent and ${}_2F_2$ the generalized hypergeometric function. ${}_2F_2$ is defined as

$${}_2F_2((a_1,a_2),(b_1,b_2),c) := \sum_{k=0}^{\infty}\frac{c^k}{k!}\frac{(a_1)_k(a_2)_k}{(b_1)_k(b_2)_k}\,, \tag{1.37}$$

see e.g. Erdelyi (1953).

For analysis of the entire system we carried out numerical simulations according to the multinomial sampling scheme. Starting from an initially homogeneous population of master individuals, we allowed for equilibration for 10^4 generations and then averaged the frequencies over another

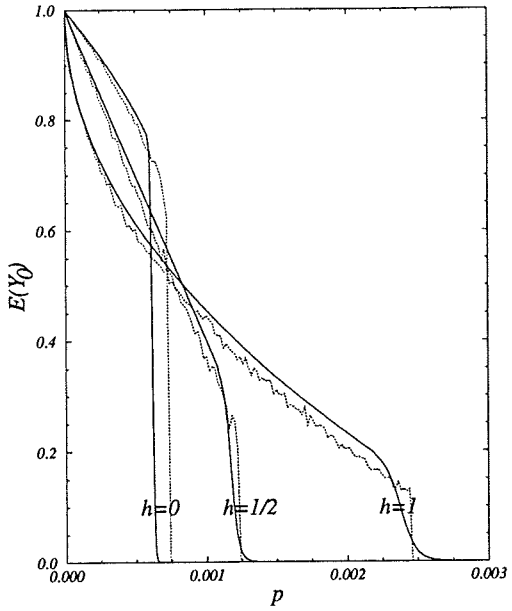

Figure 1.7: Parameters: $\nu = 30$; $s = 1/20$; $h = 0, 1/2, 1$; $N = 10^3$. Dotted: Simulation results. Solid: $E(Y_0)$ according to expressions (1.34) to (1.36). Differences between simulation and analytical results are larger for small populations; they are due to the simplifying assumption for the analytical model that the population occupies non–master error classes uniformly. In reality, however, there is a bias towards lower Hamming classes due to inflow from the master class. This effect is more drastic for small populations.

$2 \cdot 10^3$ generations. We would like to mention that the averages obtained in this way need not be estimates of frequency expectation in the strict sense: Although, if $p > 0$, the state space of the underlying Markov chain is irreducible without absorbing states (i.e. every state is recurrent), single trajectories need not be self-averaging on the time scale of a simulation, neither on evolutionary time scales. Consequently, dependence on initial conditions is observed under certain circumstances. For $h = 0$ and starting with a 'worst case initial condition' (see above), the master is usually not recovered. The existence of such metastable states is the stochastic analogue of the deterministic multiple stability which we observed for the differential equation.

In Figure 1.7 simulation results are compared with the approximate results from Eq.s (1.34) to (1.36). To determine thresholds, we numerically evaluate $p = p_{max}$ s.t. $E(Y_0) \leq 0.01$. See Table 1.1.

We investigated the effect of stochasticity by a second approach and modeled the replication process by means of a time homogeneous birth-death process as suggested by Gillespie (1976). Results are in good qualitative agreement with those of the diffusion model. The birth-death process is not subject to restrictions on the magnitude of involved parameters. Nevertheless, we prefer here to present the diffusion model, since it generalizes more easily to the two-locus situation, which will be our concern in Chapter 4.

Parameters				Simulation/Numerical	Analytical	Relative shift
ν	N	s	h			
10	100	0.05	0.5	$2.7\ 10^{-3}$	$4.3\ 10^{-3}$	56
10	∞	0.05	0.5	$6.1\ 10^{-3}$	$4.9\ 10^{-3}$	/
30	100	0.05	0	$3.9\ 10^{-4}$	$4.1\ 10^{-4}$	52
30	100	0.05	0.5	$4.6\ 10^{-4}$	$6.0\ 10^{-4}$	73
30	100	0.05	1	$5.2\ 10^{-4}$	$1.1\ 10^{-3}$	84
30	1000	0.05	0	$7.4\ 10^{-4}$	$6.5\ 10^{-4}$	10
30	1000	0.05	0.5	$1.3\ 10^{-3}$	$1.3\ 10^{-3}$	24
30	1000	0.05	1	$2.5\ 10^{-3}$	$2.6\ 10^{-3}$	22
30	10000	0.05	0	$8.2\ 10^{-4}$	$6.9\ 10^{-4}$	0
30	10000	0.05	0.5	$1.6\ 10^{-3}$	$1.6\ 10^{-3}$	6
30	10000	0.05	1	$3.1\ 10^{-3}$	$3.2\ 10^{-3}$	3
30	∞	0.05	0	$8.2\ 10^{-4}$	$8.2\ 10^{-4*}$	/
30	∞	0.05	0.5	$1.7\ 10^{-3}$	$1.6\ 10^{-3}$	/
30	∞	0.05	1	$3.2\ 10^{-3}$	$3.2\ 10^{-3}$	/
100	100	0.05	0.5	$1.2\ 10^{-4}$	$1.8\ 10^{-4}$	76
100	1000	0.05	0.5	$3.6\ 10^{-4}$	$3.8\ 10^{-4}$	27
100	10000	0.05	0.5	$4.7\ 10^{-4}$	$4.7\ 10^{-4}$	4
100	∞	0.05	0.5	$4.9\ 10^{-4}$	$4.9\ 10^{-4}$	/
30	100	0.5	0.5	$8.8\ 10^{-3}$	**	37
30	1000	0.5	0.5	$1.2\ 10^{-2}$	**	14
30	10000	0.5	0.5	$1.3\ 10^{-2}$	**	7
30	∞	0.5	0.5	$1.4\ 10^{-2}$	$1.3\ 10^{-2}$	/

Table 1.1: Comparison of threshold values for different parameters. Initial population consists exclusively of master alleles. First column shows simulation results (N finite) and results by numerical integration (N infinite). Analytical results are evaluated according to formulas (1.34) to (1.36) (N finite) and (1.19) (N infinite), respectively. Last column: Relative shift in percent of simulation threshold values for finite N with respect to deterministic value .

*) This value coincides with the *saddle–node* bifurcation point of the equilibrium frequency of y_0, viewed as function of p. Due to bistability in this case, formula (1.19) does not predict the correct threshold for the above initial condition.

**) The diffusion approximation is not valid in this case.

1.6 Summary

Error thresholds, viewed as a relationship between maximum genome length and maximum mutation rate, have been found to be important for the evolution of haploid populations of quickly-reproducing species with relatively small genomes, like in vitro evolution of RNA, evolution of bacteria and viruses. Evidence for some species of the latter to live close to that threshold has been reported repeatedly (Eigen & Schuster, 1977; Eigen *et al.*, 1989; Eigen, 9937).

A similar phenomenon might be relevant for higher organisms as well, where mutation rates are lower but genomes are larger both by a factor of at least 10^3. To match the situation in higher organisms additional features must be considered which concern the mutation as well as the selection mechanism.

As to the former, the assumption of mutation occurring exclusively as replication errors is probably a good approximation for organisms which replicate very quickly. This is, however, not the case for higher organisms where generation times are by several orders of magnitude longer than the time needed for DNA replication. Instead, independent mutations—due to radiation effects etc.—are expected to play an important part, but the relative proportions of both contributions are unresolved.

Our analyses of the two extreme cases in the form of the coupled and decoupled mutation selection equation reveals that the qualitative feature of the error threshold is not affected by the details of the mutation mechanism. However, the similarity for the parameter values considered should not obscure the fact that the relevant quantities for the threshold position are the ratio versus difference in reproduction rates for the coupled and decoupled equation, respectively.

As to the selection mechanism, diploidy effects are an immediate concern in higher organisms. We focussed on dominance, which is easily accessible within the framework adopted. We find remarkable shifts in equilibria due to dominance effects. A dominant advantageous mutation is able to persist in the population at mutation rates much higher than in the intermediary or even recessive case. This provides an aspect of evolution of dominance which is quite different from the dominance modifier theory suggested by Bürger (1983). Apart from these quantitative effects, diploidy introduces new qualitative features, like multiple equilibria, even in the case of the simple single-peaked landscape.

We remark that transitions from an ordered quasi-species into a randomly replicating ensemble as the nucleotide mutation probability is increased has also been found for the case of non symmetric selection matrices (Stadler *et al.*, 1994).

Our findings carry over to a stochastic description of the mutation selection process in finite populations. The shift of error thresholds to lower values of the mutation parameters is the more pronounced the smaller the population is, as was observed for a haploid selection regime before (Nowak & Schuster, 1989). However, we take a different way to diagnose error thresholds by

concentrating on analysis of expectations rather than maxima of the stationary distribution. Dominance effects are recovered which strongly resemble the deterministic picture: Error thresholds are shifted, and bistability re-emerges as metastability.

Beyond that, the examples listed in Table 1.1 suggest that the effects of finite population, for a given population size, saturate with sequence length.

The results arrived at so far all depend on the assumption of a single-peaked fitness landscape. We are well aware of the fact that sophisticated multi-peaked landscapes are needed to describe real-life evolution. However, apart from the fact that this 'classical' landscape allows at least some analytical treatment, it is relevant on its own right. Evolution may be conceived as a process during which advantageous mutations are introduced into a population occasionally as the result of a stochastic process. When such events occur we are – as far as the fitness landscape is concerned – in a situation close to the single-peaked one, and the new allele will only manage to establish itself if mutation rates are not too high.

2

Are there Error Thresholds for General Fitness Landscapes?

In this chapter we concentrate on the haploid version of the coupled mutation–selection model (1.1) and (1.10). Previously, we studied the phenomenon of the error threshold always assuming a certain type of fitness landscape, the single peaked landscape. We observed that relative stationary frequency of the wild type monotonically decreases as nucleotide mutation probability is increased. To numerically identify the error threshold we determined the maximal mutation probability beyond which the wild type (the master) has practically vanished from the population. In the deterministic setting this meant that its frequency would be approximately equal to $\epsilon = \frac{1}{\kappa^\nu}$ (cf. chpt. 1). Equivalently, the threshold location can be determined from population quantities like the average Hamming distance $E(p)$ or the variance $V(p)$. The former saturates sharply at $\nu/2$ (cf. Figure 1.2), the latter shows a passage through a maximal value and a subsequent decrease to approximately $\nu/4$ (cf. Figure 1.3) when the threshold is surpassed. It has been noted widely that for single peaked landscapes (abbreviated L_{SP}) the error threshold bears analogies to phase transitions in statistical physics (Leuthäusser, 1987; Higgs, 1994). Above the threshold the population composition is (almost) uniform and further increase of mutation rate does not alter the stationary distribution anymore (cf. Figure 1.1). In this case the threshold coincides with a transition from a localized stationary distribution into a delocalized one, which is spread uniformly throughout sequence space. In addition, the order of the phase transition depends in the diploid case on the dominance parameter h.

Analogies between phase transitions and error thresholds have also been observed for landscapes different from single peaked (Tarazona, 1992).

The strategy we take here is to continue to study the system of differential equations describing the replication dynamics. In particular, we will extend the study of equilibria to different fitness

landscapes and use the quantities 'master frequency' and 'average Hamming distance' to identify error thresholds.

However, as we will see in this chapter, the whole threshold concept depends not only on the shape of the landscape but also on the question of whether the polynucleotide chain is modeled as a finite or an infinite one.

Threshold formulae – like (1.19,1.20) – may be solved for ν. In this representation the error threshold has been interpreted in an information theoretical way (Eigen & Schuster, 1977): The acquisition of information in form of a specific nucleotide sequence is – for a given mutation rate – limited by a maximal chain length. Surpassing this limit means that an information carrying – and therefore 'advantageous' – wild type allele can – due to replication inaccuracy – not be maintained in the population. This observation led to an intense debate about how nature managed to avoid this so–called 'information crisis' and how, despite of that, it had been able to create organisms with genome length orders of magnitudes larger than the most primitive ones which would just bear the potential for reproduction.

If the information threshold however is directly linked to the error threshold then it also vanishes as the latter disappears under certain circumstances.

We study a simplified version of (1.1). As done in the derivation of formulae (1.19,1.20), we neglect back-flow from higher to lower Hamming classes. Thus, mutations may only accumulate and not restore a former configuration with less mutations (with respect to the wild type). If ν is finite, then the frequency of low Hamming classes (in particular y_0) is very well approximated by this simplification. For $\nu >> 1$ and in the limiting case $\nu \to \infty$ we substitute the nucleotide mutation *probability* p by the genome mutation rate λ and the binomial mutation probabilities m_{ij} by Poisson mutation rates $\hat{m}_{ij}$ (see matrices M and $\hat{M}$ in the following). Clearly, p and λ are related by $\lambda = \nu p$. In the appendix we show that mutation matrices for the cases with and without back-flow are entry-wise asymptotically equivalent as ν becomes large.

We denote the fitness values for Hamming classes (=mutation classes) by v_i ($i \in \{0, ..., \nu\}$), and mean fitness $\sum_{i=0}^{\nu} y_i v_i$ by $\bar{v}$.

We compare relevant quantities for the single peaked landscape L_{SP} having the spike $v_0 = 1$ and fitness $v_i = 1 - s$ elsewhere with the smooth landscape $v_i = (1 - s)^i$, denoted by L_H. The latter is the classical multiplicative landscape, studied for example by Haigh (1978). Later on, we superimpose both and give a threshold formula for such more general landscapes.

Finally, we generalize the multiplicative landscape within the concept of epistasis and derive threshold formulae for this case as well.

All fitness landscapes which we study in this chapter fulfill the homogeneity condition from Chapter 1.

2.1 Spiky versus Smooth

The last remark justifies the use of the mutation–selection equation in its reduced form for Hamming classes

$$\dot{y}_k = \sum_{i=0}^{\nu} y_i v_i m_{ki} - y_k \bar{v}\,, \quad (k = 0, ..., \nu)\,. \tag{2.1}$$

The entries of the $(\nu+1) \times (\nu+1)$ mutation matrix M are

$$m_{ij} = \begin{cases} \binom{\nu-j}{i-j} p^{i-j} (1-p)^{\nu-i}\,, & \text{if } i \geq j \\ 0\,, & \text{if } i < j\,. \end{cases} \tag{2.2}$$

For the Poisson approximation for large ν matrix M is replaced by $\hat{M}$ with

$$\hat{m}_{ij} = \begin{cases} \exp^{-\lambda} \frac{\lambda^{i-j}}{(i-j)!}\,, & \text{if } i \geq j \\ 0\,, & \text{if } i < j\,. \end{cases} \tag{2.3}$$

The fact that one works without back-flow mutations renders an interpretation of p not as a *nucleotide* but as *genic* mutation probability more natural. We may think of a whole segment of the DNA (instead of a single nucleotide) to mutate with probability p into a non–wild type form. The longer the segment the more unlikely is back-mutation to the wild type form and, strictly speaking, only then neglect of back-mutations in the model is justified.

We obtain the following analytical expressions of stationary frequencies and expectations of the stationary distribution for the two landscapes L_{SP} and L_H (y_i are readily proven by straightforward calculation to satisfy $\sum_{j=0}^{i} y_j v_j m_{ij} = y_i \bar{v}$; see Appendix).

Mutation matrix	Landscape L_{SP}	Landscape L_H
M	$y_0 = \frac{(1-p)^\nu - (1-s)}{s}$	$y_i = \binom{\nu}{i}(p/s)^i(1-p/s)^{\nu-i}$
	$E(p) = \frac{\nu p (1-p)^{\nu-1}}{(1-p)^{\nu-1} - (1-s)}$	$E(p) = \frac{\nu p}{s}$
$\hat{M}$	$y_0 = \frac{\exp^{-\lambda} - (1-s)}{s}$	$y_i = \exp^{-(\lambda/s)} \frac{(\lambda/s)^i}{i!}$
	$E(\lambda) = \frac{\lambda \exp^{-\lambda}}{\exp^{-\lambda} - (1-s)}$	$E(\lambda) = \frac{\lambda}{s}$

For L_{SP} frequencies $y_i\,, i > 0$ may be obtained recursively from the relation $\sum_{j=0}^{i} y_j v_j m_{ij} = y_i((1-s)+sy_0)$; an easy representation is not available in this case. We show in the Appendix how y_0 and $E(p)$ are derived. However, more importantly, in both cases we observe that the condition

$$y_0 = 0$$

is equivalent to

$$y_i = 0\,, \text{ if } i < \nu \text{ and } y_\nu = 1$$

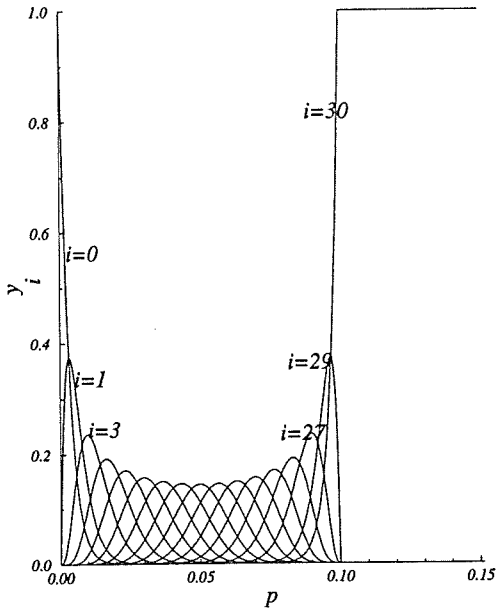

Figure 2.1: Relative stationary frequencies of Hamming classes in dependence of p. Multiplicative landscape, i.e. $v_i = (1-s)^i$. Parameters: $s = 0.1$, $\nu = 30$. Threshold at $p_{max} = s = 0.1$.

for finite ν and to

$$y_i = 0 \text{ for all } i$$

for ν infinite. For L_{SP} this property may easily be seen from the fact that the recursion relation together with $y_0 = 0$ implies $y_i = 0$, except for y_ν (if ν finite).

The error threshold is straightforwardly identified as that p-value which leads to a total loss of the master allele. At the same time, this condition coincides with population traits: The condition of a vanishing master already determines the whole distribution. As threshold condition one may therefore equivalently formulate

$$\begin{aligned} E(p) &= \nu\,, \\ E(\lambda) &= \infty\,, \end{aligned} \tag{2.4}$$

respectively.

Applying these criteria we do not detect a threshold in case L_H–$\hat{M}$. In fact, extrapolating the case L_H–M would yield $\lambda = \nu \cdot s$, which means $\lambda = \infty$, if $s \neq 0$ and $\lambda = 0$, if $s = 0$. This observation fits into the picture we have from the stationary distribution for the multiplicative landscape. The function $E(\lambda)$ has a singularity for finite λ in case L_{SP}, whereas it linearly increases in case L_H.

Our criteria for threshold detection lead to the following results

Mutation matrix	Landscape L_{SP}	Landscape L_H
M	$p_{max} = 1-(1-s)^{(1/\nu)}$	$p_{max} = s$
$\hat{M}$	$\lambda_{max} = -\log(1-s)$	$\lambda_{max} = \infty$

In fact, these threshold values satisfy simultaneously $y_0(p_{max}) = 0$ and $E(p_{max}) = \nu$ (for λ respectively). The result for L_{SP}–M differs from formula (1.20) only because of the scaling $v_0 = 1$, which we applied here; previously we had $v_0 = 1 + s$. Expanding the power to first order in s the threshold has the easy representation $p_{max} \approx \frac{s}{\nu}$.

What is particularly remarkable about these results is that the threshold in case L_H–M does not depend on ν. If the fitness distribution follows the multiplicative landscape then the same minimal replication accuracy is needed in order not to loose the master allele from the population, no matter how large polynucleotide strings are. In this sense, the 'information crisis' vanishes as the landscape L_{SP} is substituted by L_H.

2.2 Infinitely Large ν. Fitness Strictly Positive versus Truncation Selection

In the limiting case of large ν a single quantity such as y_0 becomes less robust for numerical threshold detection. For that reason it is much more reliable to switch from the study of y_0 to a study of the behavior of a population statistic – such as $E(\lambda)$ or $V(\lambda)$. To establish a relationship between arbitrary landscapes and such phenomena as error thresholds does not appear to be feasible at this stage of our research. In the following we present some case studies.

Let L_δ^Δ be defined by

$$\infty > \Delta \geq v_i \geq \delta > 0 \text{ for all } i \tag{2.5}$$

and $L_{\hat{\nu}}$ by

$$v_i = \begin{cases} \text{arbitrary} & \text{for } i \leq \hat{\nu} \\ = 0 & \text{for } i > \hat{\nu}\,, \end{cases} \tag{2.6}$$

where $\hat{\nu} \leq \nu$. L_δ^Δ contains as a special case the single peaked landscape L_{SP}.

Genes which carry too many mutations with respect to the wild type may become lethal. This idea is captured in the definition of $L_{\hat{\nu}}$. Landscapes of this type have also been associated with the notion of 'truncation selection' (Kondrachov & Crow, 1991). They conjecture that truncation selection may be present in density regulated populations, whereas weak or no epistasis (L_{SP} may be seen as a fitness landscape where epistatic effects are absent, cf. section 2.4) may be found more likely in exponentially growing populations. With this hypothesis, truncation selection might be more relevant for higher organisms, whereas landscapes without epistatic effects might be more suitable for populations of bacteria, RNA's or within the framework of prebiotic evolution.

At equilibrium, we have

$$\sum_{i=0}^{k} y_i v_i \hat{m}_{ki} = y_k \bar{v} \tag{2.7}$$

and thus

$$\bar{v}E(\lambda) = \bar{v}\sum_k k y_k =$$

$$\sum_k \sum_{i=0}^{k} i y_i v_i \exp^{-\lambda} \frac{\lambda^{k-i}}{(k-i)!} + \sum_k \sum_{i=0}^{k} (k-i) y_i v_i \exp^{-\lambda} \frac{\lambda^{k-i}}{(k-i)!} . \tag{2.8}$$

The two series in (2.8) are products of series each. The first one simplifies to

$$\exp^{-\lambda} \sum_k k y_k v_k \cdot \sum_k \frac{\lambda^k}{k!} = \sum_k k y_k v_k ,$$

the second one to

$$\lambda \exp^{-\lambda} \sum_k \sum_{i=0}^{k-1} y_i v_i \frac{\lambda^{k-1-i}}{(k-1-i)!} = \lambda \bar{v} \sum_k y_{k-1} .$$

Thus, we have

$$\bar{v}E(\lambda) = \lambda \bar{v} + \sum_k k y_k v_k . \tag{2.9}$$

For L_δ^Δ a lower bound for the last series is trivially given by

$$\sum_k k y_k v_k \geq \delta E(\lambda) .$$

This yields

$$E(\lambda) \geq \frac{\lambda \bar{v}}{\bar{v} - \delta} \geq \frac{\lambda \delta \exp^{-\lambda}}{\Delta \exp^{-\lambda} - \delta} . \tag{2.10}$$

The last inequality can be seen as follows. Let $\hat{k}$ be (for a fixed λ) the lowest index such that $y_{\hat{k}} \neq 0$. Then $y_{\hat{k}} v_{\hat{k}} \hat{m}_{\hat{k}\hat{k}} = y_{\hat{k}} \bar{v}$. Thus, $\bar{v} = \exp^{-\lambda} v_{\hat{k}} \geq \exp^{-\lambda} \delta$ and $\bar{v} \leq \exp^{-\lambda} \Delta$, independent of $\hat{k}$. Clearly, the last expression shows a singularity for

$$\lambda = -\log(\frac{\delta}{\Delta}) , \tag{2.11}$$

which is an upper bound for an error threshold for any landscape of type L_δ^Δ:

$$\lambda_{max} \leq -\log(\frac{\delta}{\Delta}) . \tag{2.12}$$

Although this is a very rough estimate, it shows the important fact that some finite λ yields $E(\lambda) = \infty$.

On the other hand, for L_ν we rewrite (2.9) as

$$E(\lambda) - \lambda = \frac{\sum_{i=0}^{\infty} i y_i v_i}{\sum_{i=0}^{\infty} y_i v_i} = \frac{\sum_{i=0}^{\nu} i y_i v_i}{\sum_{i=0}^{\nu} y_i v_i} . \tag{2.13}$$

If there was a threshold for this landscape, then the left hand side would become infinitely large for some finite λ_{max}. Since both sums on the right hand side are finite, the only way to avoid a contradiction is that the denominator becomes *zero*. That meant that all viable individua are extinct ($y_i = 0$ for all i with $v_i \neq 0$), which implies that all are extinct at equilibrium; this contradicts the condition $\sum y_i = 1$. Thus $E(\lambda)$ has to be finite for finite λ, which means that there is no threshold (in the above sense). It is easy to derive an upper bound for $E(\lambda)$ in this situation. We consider the 'worst case' $v_i = 0$ for all $i \neq \hat{\nu}$. Immediately follows $\bar{v} = v_{\hat{\nu}} y_{\hat{\nu}}$ and from (2.7) for $k \geq \hat{\nu}$

$$y_k = \exp^{-\lambda} \frac{\lambda^{k-\hat{\nu}}}{(k-\hat{\nu})!} \,. \tag{2.14}$$

The latter is independent of the particular choice for $v_{\hat{\nu}}$. Of course, for $k < \hat{\nu}$, one has $y_k = 0$. Summing over k we derive

$$E(\lambda) = \sum_k k y_k = \lambda + \hat{\nu} \,. \tag{2.15}$$

We see that $E(\lambda)$ is limited above by an affine linear function and can thus not exhibit a singularity for finite λ.

Similar results for a slightly different model and a time discrete formulation of the mutation–selection equation instead of a continuous one have been observed by Wagner & Krall (1993). Their model includes only first–order mutations (i.e. mutations from class i to $i+1$). With this assumption they showed in particular that a threshold – in the sense that the master allele gets lost – exists only if there is a strictly positive lower bound to the fitness values $(v_i)_i$, where the sequence $(v_i)_i$ is monotonously decreasing. If the lower bound is identical *zero* then no threshold is to detect.

2.3 Superimposed: Spiky and Smooth

L_{SP} and L_H both belong to a class of 'non-rugged' landscapes, in which the accumulation of mutations is deleterious and the only 'summit' of the landscape lies at v_0. They can be viewed as special cases of the following more general landscape L defined by

$$v_i = (1-\gamma)(1-s)^i + \gamma \,. \tag{2.16}$$

Clearly, the extreme cases are

$$\begin{aligned} s = 0 &: \text{ Neutral landscape}, \\ \gamma = 0 &: \text{ Multiplicative landscape}, \\ s = 1 &: \text{ Single peaked landscape}, \\ \gamma = 1 &: \text{ Neutral landscape}. \end{aligned}$$

By 'neutral' we mean that all fitness values are identical, therefore selection is absent. We know from our previous considerations the error thresholds at the boundaries. For the case of finite ν and mutation according to M we have

$$\begin{aligned} s = 0 &: p_{max} = 0\,, \\ \gamma = 0 &: p_{max} = s\,, \\ s = 1 &: p_{max} = 1 - \gamma^{1/\nu}\,, \\ \gamma = 1 &: p_{max} = 0\,, \end{aligned}$$

and for $\nu >> 1$ and mutation according to $\hat{M}$ we have

$$\begin{aligned} s = 0 &: \lambda_{max} = 0\,, \\ \gamma = 0\,,\ s \neq 0 &: \lambda_{max} = \infty\,, \\ s = 1 &: \lambda_{max} = -\log(\gamma)\,, \\ \gamma = 1 &: \lambda_{max} = 0\,. \end{aligned}$$

In fact, multiplying the four equations for case M on both sides by ν and taking the limit $\nu \to \infty$ yields exactly the four equations for case $\hat{M}$.

Although L may be viewed as a simple superposition of L_H and the uniform landscape L_1^1 ($v_i = 1$ for all i), a similar superposition of the according equilibrium distributions does not yield the distribution of the superposition. However – working without back-flow – an analytical threshold formula for an arbitrary choice of parameter values s and γ can be obtained. At equilibrium, the master frequency satisfies

$$\begin{aligned} (1-p)^\nu &= \sum_k y_k v_k = \\ (1-\gamma)\sum_k y_k (1-s)^k + \gamma &= (1-\gamma)(1-s)^\nu + \gamma\,. \end{aligned}$$

The last equation is due to the fact that, if a threshold exists, the distribution will be concentrated in class ν when the threshold is surpassed. Thus, we have

$$p_{max} = p_{max}(s,\gamma) = 1 - \Big((1-s)^\nu(1-\gamma) + \gamma\Big)^{1/\nu}\,. \tag{2.17}$$

Expanding to first order in s and $g := \gamma^{1/\nu}$, this formula simplifies to

$$p_{max} \approx s \cdot (1 - \gamma^{1/\nu})\,, \tag{2.18}$$

which is just the linear interpolation of $p_{max}(s=0)$ and $p_{max}(s=1)$.

In Figure 2.2 contour lines of the error threshold on the $s-g$–unit square are plotted. We compared analytical formula (2.17) with a numerical evaluation of the ODE system belonging to the landscape L. To obtain the numerical threshold we look for p such that $E(p) = E(p,\gamma,s)$

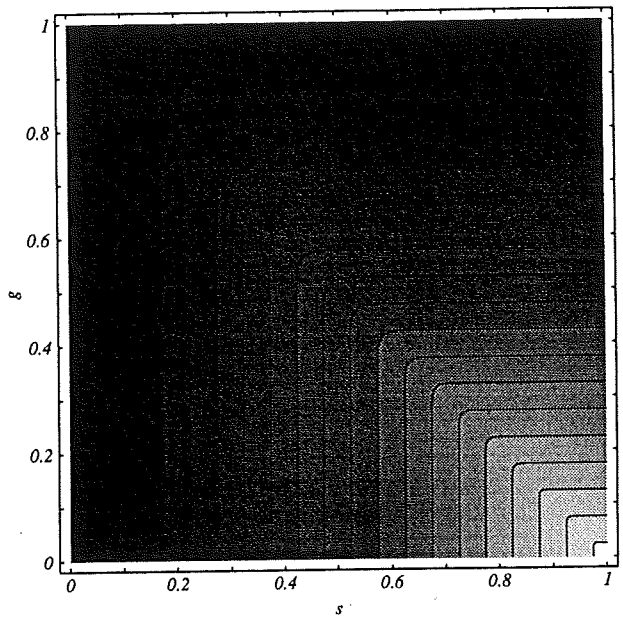

Figure 2.2: Contour lines of threshold for landscape L evaluated by analytical formula (2.17). Vertical axis is transformed according to $g := \gamma^{1/\nu}$. Parameter: $\nu = 30$. White: $p_{max} = 1$, Black: $p_{max} = 0$. Curves show a discrete sample of contour lines.

saturates and double check the result by the condition $V(p)=V(p,\gamma,s) \leq \epsilon$. The numerical method of choice is determination of the stationary distribution as the eigenvector belonging to the largest eigenvalue of the product matrix $M \cdot (diag(v))$ (for the haploid system the stationary distribution is unique; see remarks on page 14). Here, $(diag(v))$ means the diagonal matrix with entries v_i on its main diagonal and 0's elsewhere. Evaluating eigenvectors is computationally much faster than numerical integration of the *ODE* system. We used a standard Fortran software routine (Smith *et al.*, 1976). On a grid of the unit interval with equidistant spacing $\Delta s = \Delta g = 0.1$ we observed a maximal discrepancy of numerical and analytical threshold values of less than 0.5%.

By our former remarks on landscape L_δ^Δ it is clear that for the case of infinite ν landscapes of the form L still lead to a finite threshold for any combination of parameters s and γ, except for the degenerate case $\gamma = 0$, where $\lambda_{max} = \infty$. The threshold formula for the case of infinite ν becomes $\lambda_{max} = -\log(\gamma)$, independent of s, if $s \neq 0$. Of course, $s = 0$ still means neutrality, hence $\lambda_{max}(s = 0) = 0$.

The function $p_{max}(s, g)$ shows a remarkable symmetry. Threshold determining is either exclusively the parameter g or the parameter s depending on whether $s \geq (1 - g)$ or $s < (1 - g)$. Concavity of the contour lines implies that a linear superposition of any two landscapes of type L, characterized by parameters (s_1, g_1) and (s_2, g_2) may only increase the threshold. For instance, considering $(s_1, g_1) = (s, 0)$ and $(s_2, g_2) = (0, 1-s)$, one deduces for both of them the same threshold $p_{max} = s$. A convex combination of these two landscapes may increase the threshold level to

up to $p_{max} \approx \frac{1+s}{2}$. That means that a suitable combination of fitness functions may on the spot enhance error tolerance of the replication process. It is conceivable that such transformations of landscapes, in other words a 'changing environment', can be realized by means of time dependence of the fitness function.

Think of some finite population of sequences of finite length. At an early stage of evolution the landscape might resemble very much a single peaked one, when some particular constellation has a definite fitness advantage. As evolution proceeds and the population gets better adapted, the center of the quasi-species shifts in direction of the peak. This process is like a zoom of the landscape. As the population is more concentrated around the peak, the fine structure of the landscape close to the peak becomes more 'visible'. This could be interpreted as being equivalent to a move on a contour line from initially $(1, g)$ in direction (g, g). If better adaptation can be captured in terms of a smoothened landscape, then a further move on the contour line towards $(g, 0)$ may be hypothesized. These moves have the property that at no stage a more accurate replication mechanism was required, since error thresholds stay constant. However, as the initial single peaked landscape transformed into a multiplicative one, the dependency of replication accuracy on chain length vanished, which implies, that restrictions on maximal chain length – formerly present – are removed. Chain length may now be increased – for example to encode a better replication system. This could create individua with a definite fitness advantage and return the landscape into single peaked.

In Figure 2.3 we show how an initially single peaked landscape transforms into a multiplicative one as one moves on contour lines.

All examples of landscapes considered so far may be contested on reasons of being too unrealistic for 'natural' fitness landscapes. More realistic ones, for example for tRNA's, may be constructed by observing sequence–secondary structure relationships. Nonetheless, we claim the above quoted examples to be conceptually important.

For instance, in the context of tRNA's secondary structures may be viewed as a *phenotype* belonging to a particular sequence, its genotype. Dynamics and threshold behavior at this phenotypic level has been studied by Reidys *et al.* (1994). They observe similar threshold patterns if the fitness landscape is analogously single peaked, i.e. if *one* highly fit structure (instead of sequence) competes against other structures, all of the same low fitness.

In Chapter 1 we concluded that diploidy led to threshold shifts, the shifts depending on dominance properties. However, it did not cause thresholds to disappear, as can be observed here e.g. for a multiplicative landscape as ν becomes large. How does the dynamics look for the diploid case when fitness values are multiplicative? We apply model (1.1) to the fitness assignments $w_{ij} = (1-s)^{i+j}$, which is the most obvious transfer of the haploid multiplicative model to the diploid situation. The dynamics of the Hamming class frequencies y_i is then up to time transformation identical to the one of the haploid situation. In particular, equilibria remain

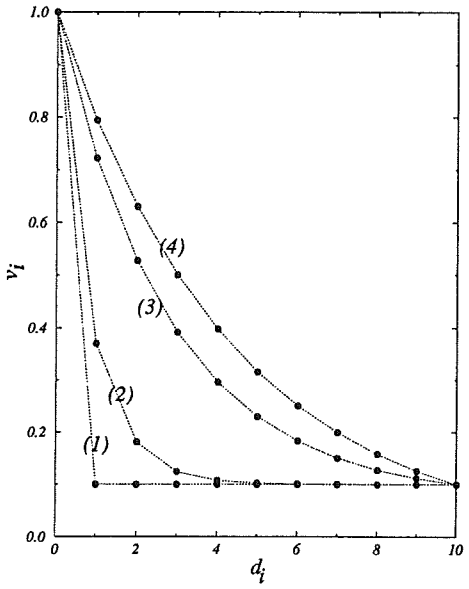

Figure 2.3: Fitness landscapes. Abscissa: Hamming distance $d_i := d(A_0, A_i)$ with respect to the master sequence. Ordinate: Fitness value $v_i = v_i(s, \gamma)$ as defined in (2.16). Shown are four different landscapes which yield the same threshold value $p_{max} = 0.206$ as one moves along the contour line starting from L_{SP} with $(s, \gamma) = (1.000, 0.100)$ (plot (1)) to L_H with $(s, \gamma) = (0.206, 0.000)$ (plot (4)). Intermediary are $(s, \gamma) = (0.700, 0.100)$ and $(s, \gamma) = (0.300, 0.074)$ (plots (2) and (3)). Parameter: $\nu = 10$.

the same. This is due to the fact, that marginal fitnesses $w_i = \sum_j w_{ij} y_j$ simplify to $(1-s)^i \bar{v}$ and mean fitness to $\bar{v}^2$, where $\bar{v} = \sum_i y_i (1-s)^i$. Thus, threshold properties for L_H carry over directly from haploid to diploid dynamics. More generally, this is true whenever $w_{ij} = v_i v_j$.

2.4 Epistasis

Usually the concept of epistasis is discussed within the framework of non-additive or non-multiplicative gene interaction.

We introduce fitness functions $L_{K(\alpha)}$ with positive parameter α. They are commonly used to model epistatic fitness interactions (Kondrachov, 1982; Chao, 1988). We define

$$v_i = (1-s)^{i^\alpha} \tag{2.19}$$

and classify

$$\begin{aligned} 0 < \alpha < 1 &: \text{diminishing epistasis} \\ \alpha = 1 &: \text{multiplicativity} \\ 1 < \alpha &: \text{synergistic epistasis}. \end{aligned}$$

We emphasize the special case $\alpha \to 0$, which generates a single peaked landscape. Thus, $L_{K(\alpha)}$ generalizes the two standard landscapes L_{SP} and L_H considered before and contains them as special cases. Unfortunately though, there is no analytical solution for the stationary frequencies y_i in closed representation available for $L_{K(\alpha)}$. However, we are able to present a threshold formula also for this type of landscape.

So far, for threshold detection the two conditions 'loss of wild type' and 'saturation of $E(p)$' could be used equivalently.

This is still true for the generalized landscape $L_{K(\alpha)}$ as long as $\alpha \leq 1$. Diminishing epistasis means that the fraction $\frac{v_i}{v_{i-1}} > 1-s$, i.e. any additional mutation deteriorates fitness less than geometrically. Thus, it becomes clear that a given nucleotide mutation probability p_0 which can cause Hamming class i to disappear from the population will also cause any higher Hamming class $i+k$ to disappear. In other words maximal mutation pressure is needed to remove the wild type from the population. As soon as this is accomplished the stationary distribution is immediately concentrated in Hamming class ν.

For synergistic epistasis ($\alpha > 1$) the situation is reversed. Synergism means $\frac{v_i}{v_{i-1}} < 1-s$, i.e. any additional mutation deteriorates fitness more than geometrically. To remove class i from the stationary distribution requires stronger mutation pressure than was needed for removal of class $i-1$. Selection of individua belonging to class $i-1$ among the rest of the population when all error classes up to $i-2$ are already lost, is relatively stronger than that of those belonging to class i among the rest, when classes up to $i-1$ are lost. It implies, that there is no equivalence

of the threshold detection criteria in this situation. In fact, we have to distinguish between p_{max}, which leads to loss of the wild-type allele, and the higher mutation probability p^{max}, which leads to a saturation of $E(p)$ (or, equivalently $y_\nu = 1$). We give analytical formulae for both maximal mutation probabilities.

To derive p^{max} we calculate p such that the second to the last mutation class gets just lost at stationarity by mutation pressure. We assume that any lower mutation class ($i = 0$ to $i = \nu - 2$) is already lost. Thus, we have to satisfy the following three algebraic equations

$$\begin{aligned} y_{\nu-1}v_{\nu-1}p + y_\nu v_\nu &= y_\nu(y_{\nu-1}v_{\nu-1} + y_\nu v_\nu) \\ y_{\nu-1}v_{\nu-1}(1-p) &= y_{\nu-1}(y_{\nu-1}v_{\nu-1} + y_\nu v_\nu) \\ y_{\nu-1} + y_\nu &= 1\,. \end{aligned}$$

The solution is

$$p = \frac{((1-s)^{(\nu-1)^\alpha} - (1-s)^{\nu^\alpha})(1-y_{\nu-1})}{(1-s)^{(\nu-1)^\alpha}}\,, \tag{2.20}$$

which, for $y_{\nu-1} = 0$, yields

$$p^{max} = 1 - (1-s)^{\nu^\alpha - (\nu-1)^\alpha}\,,\ (\alpha \geq 1)\,. \tag{2.21}$$

On the other hand, the wild type is already lost if

$$p_{max} = s\,,\ (\alpha \geq 1)\,. \tag{2.22}$$

This holds for any $\alpha \geq 1$ and independently of ν. To justify the last equation we note that it holds for $\alpha = 1$ (see above). Letting α go to ∞ we obtain fitnesses $v_0 = 1$, $v_1 = 1 - s$ and $v_i = 0$ ($i \neq 0, 1$). At equilibrium y_0 satisfies $(1-p)^\nu = y_0 + y_1(1-s)$. Putting $y_0 = 0$ the latter implies $p = 1 - (y_1(1-s))^{1/\nu}$. Furthermore, y_1 then satisfies $y_1(1-s)(1-p)^{\nu-1} = y_1^2(1-s)$. Thus, $y_1 = (1-p)^{\nu-1}$. Inserting y_1 into the equation for p yields $(1-p)^\nu = (1-p)^{\nu-1}(1-s)$, thus $p = s$. Since for landscapes $L_{K(\alpha)}$ the relation $\alpha_1 < \alpha_2$ implies $y_{0_{(\alpha_1)}} \leq y_{0_{(\alpha_2)}}$ (as functions of p) we conclude that by continuity

$$0 = y_{0_{(\alpha=1)}}(p=s) \leq y_{0_{(\alpha)}}(p=s) \leq y_{0_{(\alpha=\infty)}}(p=s) = 0\,,$$

which means $y_{0_{(\alpha)}}(p=s) = 0$ for all $1 \leq \alpha \leq \infty$.

As expected, p_{max} and p^{max} coincide for $\alpha = 1$, since the two detection criteria are equivalent for L_H.

We treat the case of diminishing epistasis now. To derive the threshold formula, we make use of the fact that the two detection criteria are equivalent. Thus, we may equate

$$(1-p)^\nu = (1-s)^{\nu^\alpha}\,. \tag{2.23}$$

Solving for p leads to

$$p_{max} = 1 - (1-s)^{\nu^{\alpha-1}}\,,\ (\alpha \leq 1)\,. \tag{2.24}$$

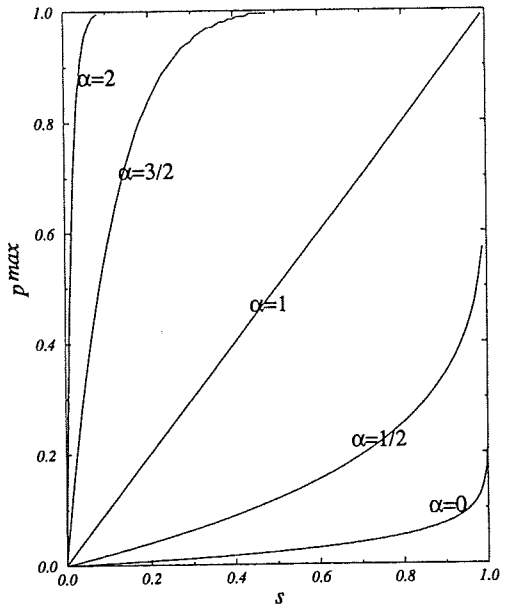

Figure 2.4: Error threshold p^{max} for generalized landscapes $L_{K(\alpha)}$. Fitness function $v_i = (1-s)^{i^\alpha}$. Parameter: $\nu = 30$.The curve $\alpha = 0$ belongs to the single peaked landscape $v_0 = 1$, $v_i = 1 - s$.

Eq. (2.23) is derived from the condition for the master frequency $y_0 v_0 m_{00} = y_0 \bar{v}$. When the threshold is surpassed then $\bar{v} = (1-s)^{\nu^\alpha}$, since the distribution is immediately concentrated in class ν.

We checked our results numerically by the method of eigenvalues and eigenvectors. Results are presented in Figures 2.4 and 2.5. Numerical and analytical curves are not distinguishable at the chosen scale.

How are relations of s, ν and α? First, one observes that thresholds p^{max} for fixed s increase as the parameter α is increased. The lower bound is given by the threshold of the single peaked landscape ($\alpha \to 0$), the upper bound ($\alpha \to \infty$) by the one of the degenerate landscape $v_0 = 1$, $v_1 = 1 - s$ and $v_i = 0$ ($i \neq 0, 1$); in this latter case one has $p_{max} = 1$.

Furthermore, as can be seen from Figure 2.5, in contrary to all what we have observed before, for $\alpha > 1$ threshold p^{max} even increases as chain length ν is increased. The correlation between ν and p^{max} for landscapes of type $L_{K(\alpha)}$ is

$$\begin{aligned} \alpha > 1 &: \text{positive correlation} \\ \alpha = 1 &: \text{independence} \\ 0 < \alpha < 1 &: \text{negative correlation}. \end{aligned}$$

Adopting p_{max} as 'threshold' the relations are

$$\begin{aligned} \alpha \geq 1 &: \text{independence} \\ 0 < \alpha < 1 &: \text{negative correlation}. \end{aligned}$$

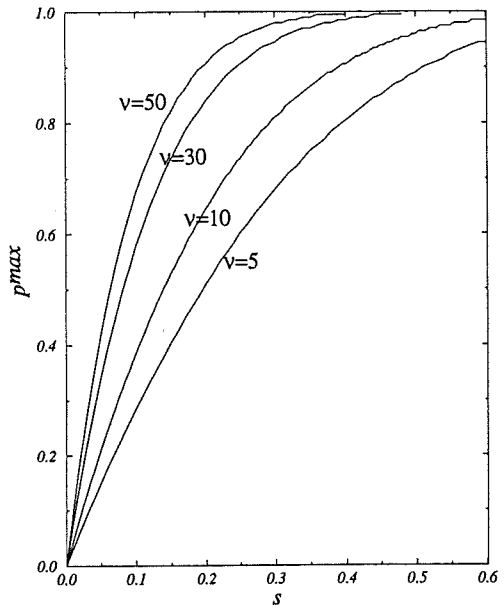

Figure 2.5: Error threshold p^{max} for generalized landscape $L_{K(3/2)}$. Curves from top to bottom represent threshold functions for ν–values 50, 30, 10 and 5. Positive correlation of ν and p^{max}. Parameter: $\alpha = 3/2$.

Thus, with equations (2.21),(2.22) and (2.24) we have generalized threshold formulae which comprise as special cases the ones for L_{SP} and L_H. In fact, choosing $\alpha = 1$ leads to $p_{max} = s$ and choosing $\alpha = 0$ to $p_{max} = 1 - (1 - s)^{1/\nu}$ (cf. results in section 2.2).

Correlation between chain length and error thresholds, as it turns out, depends in the first place on what one likes to define as 'threshold'. Using the criterion of saturation of $E(p)$ (equivalently $y_\nu = 1$), one concludes that correlation changes from negative to positive as the parameter α changes from less than 1 to bigger than 1, where independence is to observe for $\alpha = 1$. On the other hand, with a threshold definition of a vanishing wild type ($y_0 = 0$), correlation is negative for any $\alpha < 1$. For any $\alpha \geq 1$ threshold p_{max} and ν are independent.

These results, of course, obscure the fact that master frequency $y_0(p)$ follows an exponential law depending on ν: $y_0(p) = (1 - p/s)^\nu$ (take as example the case $\alpha = 1$). Thus, for finite populations (size N) the negative correlation of ν and p_{max} is still present since $(1-p/s)^\nu < 1/N$ for any choice of N if ν is sufficiently large. Both quantities become independent only in the deterministic limit.

2.5 Summary

For an easier overview we summarize the results for the various examples of fitness landscapes. All landscapes which we considered in foregoing sections are characterized by few parameters (at

	Landscape		Error Threshold	
			A) ν finite	B) ν infinite
1)	L_{SP} (single peaked)		$1-(1-s)^{1/\nu}$	$-\log(1-s)$
2)	L_H (multiplicative)		s	∞
3)	$L_{K(\alpha)}$ (epistatic)	$\alpha \leq 1$:	$1-(1-s)^{\nu^{\alpha-1}}$	∞
		$\alpha > 1$:	$\begin{cases} s\ (=p_{max}) \\ 1-(1-s)^{\nu^\alpha-(\nu-1)^\alpha}\ (=p^{max}) \end{cases}$	∞
4)	L_δ^Δ (arbitrary, bounded above and below)		$\leq 1-(\frac{\delta}{\Delta})^{1/\nu}$	$\leq -\log(\frac{\delta}{\Delta})$
5)	$L_{\hat{\nu}}$ (truncation selection)		not well defined	∞
6)	L (superimposed)		$1-((1-s)^\nu(1-\gamma)+\gamma)^{1/\nu}$	$-\log(\gamma)$

most *two*) and are therefore very 'simple' ones. We do not intend to give a complete picture but to illustrate that threshold behavior depends crucially, i.e. qualitatively, on the fitness landscape. Clearly, in case 5B, only an upper bound can be given, since landscapes L_δ^Δ are not unambiguously defined. Note, that thresholds in case A refer to p (nucleotide mutation probability), whereas in case B they refer to λ (genic mutation rate). Note further, that formula 4B comprises all other formulae in column B (as well, 4A comprises those in column A, except for the case 5A). This suggests that primarily the ratio of highest and lowest fitness values determines whether a finite error threshold can be observed or not.

In particular, for the case of infinite ν and monotonously decreasing (with respect to the Hamming class) fitness functions, we conclude that a threshold exists if and only if $\delta \neq 0$. This result coincides with that of Wagner & Krall (1993).

The study of equilibria in the deterministic case is important as a first step to acquire some insight in possible long term dynamic behavior. However, in order to draw a more realistic picture of evolutionary mechanisms on the molecular level the study of finite populations and their time dependent behavior would be necessary.

So far, we see that the problem of being restricted to ever more accurate replication in order not to loose the wild type as chain length grows is very much dependent on the chosen (or given) fitness landscape.

With our present knowledge of realistic fitness landscapes it does not seem to be clear at all to what extent the so–called information crisis and error catastrophe discussed in the sequel of the paper by Eigen (1971) present an impending problem to evolution of higher organisms.

3

Effect of Repeated Recombination

We have seen that an ordered quasi-species, i.e. a population composition centered around a master allele in sequence space, is endangered by accumulating mutations, no matter whether they are due to inaccurate replication or to environmental impact. The danger becomes manifest in two ways. A newly introduced master (i.e. an advantageous allele) may not be able to establish itself in the population if mutation pressure is too strong, but the dynamics rather approaches a suboptimal equilibrium (recessive to slightly recessive cases discussed in Chapter 1). On the other hand, an initially homogeneous population consisting exclusively of master alleles may loose this master under the influence of mutation. Both scenarios have been captured in terms of the error threshold.

The model we employed so far can be characterized as an 'n-allele-*one*-locus' model. In this chapter we will extend it to a *two*-locus situation. Thus, quite naturally, the diploid setup from Chapter 1 is augmented by the additional feature of recombination between two adjacent loci. One may think of 'loci' as being gene loci or exons with or without introns or any other arbitrary stretch of DNA in-between. The distance among the two can be captured in terms of their cross–over probability.

We will see that besides the requirement of a small enough mutation rate, the requirement to recombine sufficiently often may be essential to retain acquired information; i.e., in our terms, to retain advantageous alleles. It appears that for evolution to proceed effectively not only an error threshold has to be underscored but also some recombinational threshold might have to be overscored.

On the other hand, to break favorable combinations of chromosome stretches apart may harm an individual. Think of the following situation. Two adjacent genes may be functional only if both are present as specific alleles; function may be disrupted if one or both are present in non–wild type form. We first discuss this case in the following section and observe some influence on the error threshold.

In the second scenario the view point will be different. The basic question we are interested in is to know what happens to an equilibrated frequency distribution of the alleles present at one gene locus, if an advantageous mutation occurs at the second gene locus. The equilibrium will be distorted, the intensity being dependent on the recombination rate, while the advantageous mutation increases in frequency. However, if such events are rare the former mutation–selection balance at the linked locus may be restored. The time span depends primarily on the selective advantage of the original master and indirectly on the recombination rate. To restore the former equilibrium will be possible, if master frequency has not been shifted into the domain of attraction of another (locally stable) equilibrium. As we have seen in Chapter 1, in the diploid situation more than one stable equilibrium may exist.

Further, if the new advantageous mutation increases in frequency very quickly and the two loci are linked very tightly, the original master sequence may even go extinct due to lack of time for a recombination event to happen while the new advantageous mutation at the second locus is on its way to fixation. In this case there would be no way for the master to recover unless by back mutation.

Segmentation of the genome and to allow for recombination between the single units may be seen as a possible strategy for a molecular species to avoid a detrimental loss of information (in form of once acquired advantageous alleles). If the genome was reproduced as a homogeneous, single unit then for example a non segmented polynucleotide chain of length 10^6 kB would require an essentially exact replication in order not to loose a master sequence with a fitness advantage of any reasonable magnitude from the population.

In this chapter we deal in a first approach again with infinitely large populations and describe the process as a deterministic approximation, neglecting random forces. We come back to this point in the next chapter.

3.1 The Model

We consider two loci $\mathcal{A}$ and $\mathcal{B}$ with a finite number of nucleotides $\nu_{\mathcal{A}}$ and $\nu_{\mathcal{B}}$, respectively.

It turns out to be conceptually more convenient to recede to the decoupled version of the mutation–selection equation (cf. Eq. (1.3)) when introducing a second locus. Furthermore, we employ the caricature from section 1.5. At both loci we collect separately all Hamming classes except the master class into the error tail. Therefore, we have four distinct quantities: Masters at $\mathcal{A}$ and $\mathcal{B}$ and error tails at $\mathcal{A}$ and $\mathcal{B}$ – two loci with two distinct alleles each.

Recombination then fits in a natural way into the system and the equations become

$$\dot{x}_{ij} = (1-R)x_{ij}\tilde{w}_{ij} + R\sum_{m,n} x_{im}x_{nj}\tilde{w}_{im,nj} - x_{ij}\bar{w} + \sum_{m,n}\tilde{m}_{ij,mn}x_{mn}\,. \tag{3.1}$$

All indices in the above expression range over $\{1,2\}$. The first index in a pair belongs to the

$\mathcal{A}$ locus, the second to $\mathcal{B}$. The numbers are related to the alleles by 1 corresponding to A or B (masters) and 2 corresponding to a or b (error tails). R is the recombination fraction. x_{ij} denotes frequency of gametes carrying allele i at $\mathcal{A}$ and j at $\mathcal{B}$. $\tilde{w}_{ij,mn}$ is the replication rate (i.e. fitness) of a diploid individual with one i-j and one m-n gamete. $\tilde{w}_{ij}$ is marginal fitness of an i-j gamete and $\bar{w}$ mean fitness of the population. $\tilde{m}_{ij,mn}$ is the rate with which an m-n gamete is replicated as an i-j gamete (i.e. mutation rate).

For more than two alleles (e.g. M at $\mathcal{A}$ and N at $\mathcal{B}$) the form of this equation remains the same. Only the range of indices has to be generalized to $\{1,...,M\}$ and $\{1,...,N\}$, respectively.

Finally, here and in the following $\dot{x}$ always means differentiation with respect to time $\dot{x} = \frac{dx}{dt}$. For notational convenience we replace for the two-locus two-allele situation the occuring indices in the following way

index pair	replaced by	allelic composition
(11)	1	A–B ,
(21)	2	a–B ,
(12)	3	A–b ,
(22)	4	a–b .

Introducing the quantity $D := x_2x_3 - x_1x_4$, called *linkage disequilibrium*, and assuming the fitness matrix to be symmetric and fitness values for *cis*- and *trans*-heterozygotes to be identical, system (3.1) can be rewritten as

$$\dot{x}_i = x_i(\tilde{w}_i - \bar{w}) + \eta_i \tilde{w}_{14} RD + \sum_j \tilde{m}_{ij} x_j \,, \tag{3.2}$$

where $\eta_1 = \eta_4 = -\eta_2 = -\eta_3 = 1$.

We assume that mutations occur independently at the two loci. Furthermore, as we did before in the one–locus analysis, we neglect back-flow from a to A. Mutations are modeled as point mutations occuring independently at each nucleotide site. We do the same at locus $\mathcal{B}$. The mutation matrix can then be written as

$$\tilde{M} = \begin{pmatrix} -\mu\nu_{\mathcal{A}} - \mu\nu_{\mathcal{B}} & 0 & 0 & 0 \\ \mu\nu_{\mathcal{A}} & -\mu\nu_{\mathcal{B}} & 0 & 0 \\ \mu\nu_{\mathcal{B}} & 0 & -\mu\nu_{\mathcal{A}} & 0 \\ 0 & \mu\nu_{\mathcal{B}} & \mu\nu_{\mathcal{A}} & 0 \end{pmatrix},$$

with μ, as in Chapter 1, being nucleotide mutation rate.

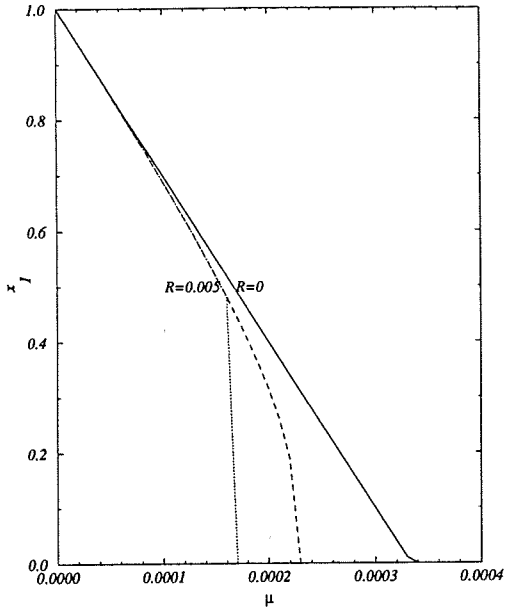

Figure 3.1: Equilibrium frequency of master A–B in dependence of mutation rate μ. Single peaked landscape model. Solid: No recombination ($R = 0$). Equilibrium is globally stable, irrespective of initial conditions. Dashed: Intermediate recombination rate $R = 0.005$ and initial condition $x_1(t = 0) = 1$. Dotted: $R = 0.005$, initial condition $x_4(t = 0) = 1$. Shift of threshold (i.e. μ such that $x_1 = 0$) depends on R and initial conditions. Parameters: $\nu_A = \nu_B = 15$; $s = 0.01$; $h = 0.5$.

3.2 Effect of Recombination on a Single Peaked Landscape

The single peaked landscape from Chapter 1 carries directly over, so that fitness values are specified as

$$\tilde{W} = \begin{pmatrix} 1+2s & 1+2hs & 1+2hs & 1+2hs \\ 1+2hs & 1 & 1 & 1 \\ 1+2hs & 1 & 1 & 1 \\ 1+2hs & 1 & 1 & 1 \end{pmatrix}, \tag{3.3}$$

with h, as before, being the dominance parameter. We observe a shift of the diploid error threshold (see Eq. (1.21)) for the single peaked landscape, which depends on the recombination rate R.

Additionally, the threshold depends on initial conditions. Recombination has a more pronounced effect if an advantageous allele is introduced ($x_4(t = 0) = 1$) than when a preexisting wild-type is endangered by mutation and recombination ($x_1(t = 0) = 1$). For analysis, we put $h = 1/2$. It is convenient to transform system (3.1) into equations for $y_A := x_1 + x_3$, $y_B := x_1 + x_2$ and x_1 (note that cis– and trans–heterozygotes have different fitness values). These read

$$\dot{y_A} = sx_1(1 - y_A) - \mu\nu_A y_A\,, \tag{3.4}$$

$$\dot{y_B} = sx_1(1 - y_B) - \mu\nu_B y_B\,, \tag{3.5}$$

$$\dot{x}_1 = sx_1(1 - x_1) + R(y_A \cdot y_B - x_1 - sx_1(1 - y_A - y_B + y_A \cdot y_B)) - \mu(\nu_A + \nu_B)x_1\,. \tag{3.6}$$

Eq. (3.6) may be simplified, if one puts $R(1 + s) \approx R$; then the term involving Rs in the

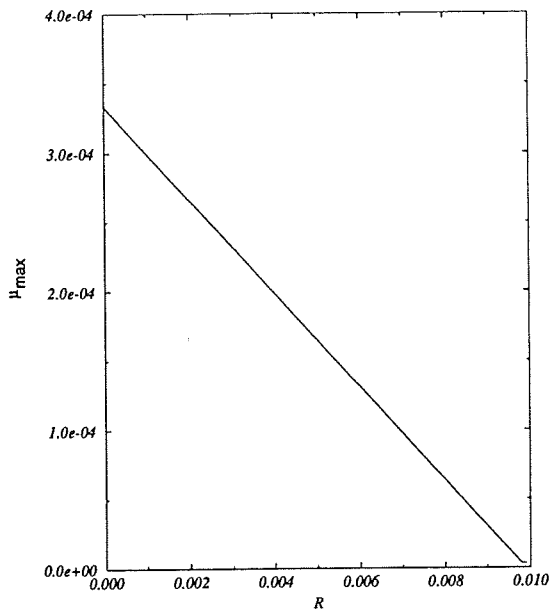

Figure 3.2: Single peaked landscape model. The error threshold μ_{max} (Eq. (3.10) for the 'worst case' $x_4(t=0)=1$) depends linearly on the recombination rate R. If $R \geq s$ advantageous gametes A–B cannot be established: $x_1(t)=0$ for all t. Parameters: $\nu_A = \nu_B = 15$; $s = 0.01$; $h = 0.5$.

above expression vanishes. The single peaked landscape means that gametes A–b and a–B behave similarly; therefore, the dynamics for y_A and y_B are essentially the same and we may put $y_A = y_B$. Steady states can now easily be detected as intersections of isoclines in the x_1–y_A phase plane. In particular, it becomes clear that, depending on parameters, multiple equilibria may exist.

To derive them explicitly is in principle possible, since the algebraic equations $\dot{x}_1 = 0$ and $\dot{y}_A = 0$ are of low enough order. Nevertheless, the expressions become very clumsy.

In addition to the error threshold a *recombination threshold* exists: The master may also be lost ($x_1(t=\infty)=0$), if recombination rate exceeds a certain limit. In Figure 3.3 we show how the equilibrium frequency x_1 decreases as the recombination rate is increased.

We present an error threshold formula for one of the possible equilibria. It fits the situation for the initial condition $x_4(t=0)=1$. The equation for x_1 has to satisfy ($\tilde{w}_{ij}$ are entries of $\tilde{W}$ and $\tilde{w}_1$ is marginal fitness of x_1)

$$0 = x_1(\tilde{w}_1 - \bar{w}) + R(x_1^2\tilde{w}_{11} + x_1x_2\tilde{w}_{12} + x_1x_3\tilde{w}_{13} + x_2x_3\tilde{w}_{23} - x_1\tilde{w}_1) + x_1\tilde{m}_{11}\,, \tag{3.7}$$

or, equivalently

$$0 = 2sx_1(1-x_1)(x_1(1-2h)+h) + R(x_2x_3 - (1+2hs)x_1x_4) - \mu(\nu_A+\nu_B)x_1\,. \tag{3.8}$$

Dividing by x_1, letting $x_1 \to 0$ and assuming that the product x_2x_3 approaches 0 more quickly than x_1 one derives

$$\mu_{max} = \frac{2hs(1-R)-R}{\nu_A+\nu_B}\,. \tag{3.9}$$

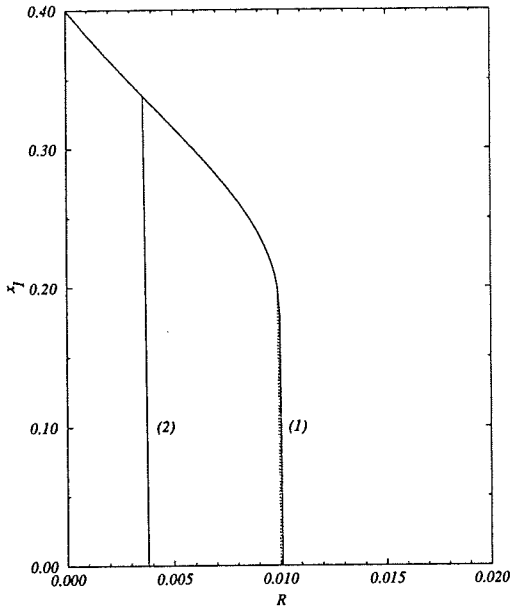

Figure 3.3: Equilibrium frequency of master A–B in dependence of recombination rate R. Single peaked landscape model. Solid: Analytical solution. Dotted: Numerical solution by integrating the *ODE* system (3.2). The master gets lost if recombination exceeds a maximal rate $R \approx s$, where initial population consists of masters only ($x_1(t=0)=1$) (1). For the initial condition 'no master' ($x_4(t=0)=1$), recombination threshold is at $R = 0.004$ (2), as predicted from formula (3.10). Parameters: $\nu_A = \nu_B = 15$; $s = 0.01$; $h = 0.5$; $\mu = 0.0002$.

Obviously, in absence of recombination ($R = 0$) the above threshold formula turns into (1.21). However, equilibria depend again on initial conditions. Formula (3.9) provides only a lower bound to the error threshold. This is due to the above assumption about the relative convergence velocities of x_1 and x_2x_3. It is justified only for the case when A–B alleles are built up, but not when they pre-exists in the population.

Eq. (3.9) may be solved for R to obtain

$$R_{max} = \frac{2hs - \mu(\nu_A + \nu_B)}{1 + 2hs} \tag{3.10}$$

as upper bound on the recombination rate beyond which the master will be lost at equilibrium.

3.3 Effect of Recombination under Directional Selection

We distinguish here two phases in the course of evolution. A 'normal' phase (phase 1) on a slow time scale, during which moderate forces of mutation and selection lead to mutation–selection balance. Such a phase is interrupted by an occasional evolutionary disturbance, when a definitely advantageous mutation arises and dominates the dynamics for a short time during which it gets established in the population (phase 2). Such advantageous mutations are assumed to have selective differentials by at least one order of magnitude higher than those of the 'normal' selected

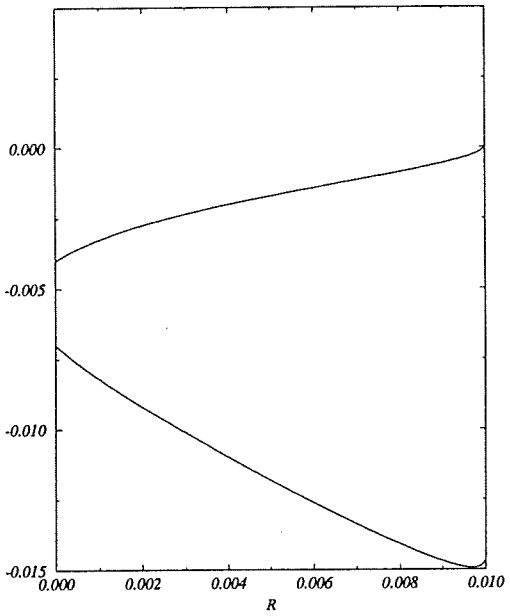

Figure 3.4: Single peaked landscape model. The two eigenvalues of the Jacobian belonging to equations (3.4) and (3.6). Both are negative as long as $R < R_{max}$, therefore the equilibrium shown in Figure 3.3 is stable. At the critical recombination rate stability is lost. Eigenvalues become complex, one of them having a positive real part. Parameters: $\nu_A = \nu_B = 15$; $s = 0.01$; $h = 0.5$; $\mu = 0.0002$.

alleles. Evolution in these fast time scale periods is dominated by selection compared to which mutation is weak.

We have to specify fitness values. We concentrate on the case of directional selection with no dominance ($h = 1/2$) at both loci. Under this condition there is only one globally stable equilibrium for the mutation–selection dynamics at each locus. For an additive fitness distribution between and within both loci matrix $\tilde{W}$ then takes the form

$$\tilde{W} = \begin{pmatrix} 1+2s_1+2s_2 & 1+s_1+2s_2 & 1+2s_1+s_2 & 1+s_1+s_2 \\ 1+s_1+2s_2 & 1+2s_2 & 1+s_1+s_2 & 1+s_2 \\ 1+2s_1+s_2 & 1+s_1+s_2 & 1+2s_1 & 1+s_1 \\ 1+s_1+s_2 & 1+s_2 & 1+s_1 & 1 \end{pmatrix} . \tag{3.11}$$

Since we are interested mainly in the relative frequency of the master allele A at $\mathcal{A}$, it turns out to be convenient to study the quantities $y_B := x_1 + x_2$, $y_A := x_1 + x_3$ and D. Then, system (3.2) transforms equivalently to

$$\dot{y_A} = s_1 y_A(1-y_A) - s_2 D - \mu\nu_A y_A , \tag{3.12}$$

$$\dot{y_B} = s_2 y_B(1-y_B) - s_1 D - \mu\nu_B y_B , \tag{3.13}$$

$$\dot{D} = D\Big(s_1(1-2y_A) + s_2(1-2y_B) - (1+s_1+s_2)R - \mu\nu_A - \mu\nu_B\Big) . \tag{3.14}$$

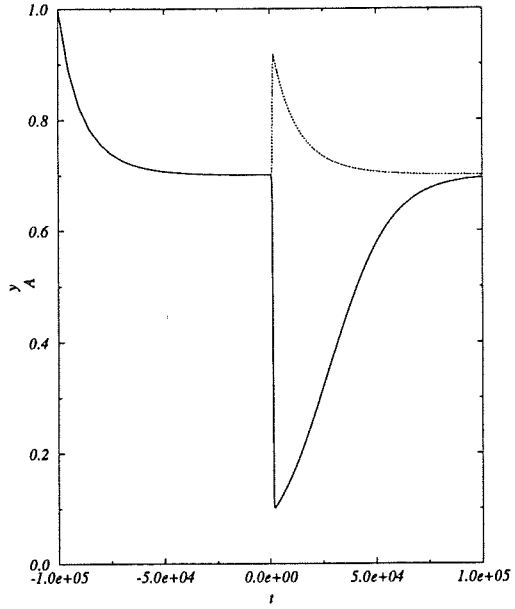

Figure 3.5: Relative frequency of master A in dependence of time t. Directional selection model. Dotted line: Mutation B arises at $t_1 = 0$ on a gamete carrying A (upper peak). Solid line: Mutation B arises at $t_1 = 0$ on a gamete carrying one of the error tail alleles a (lower peak). Phase 2 is extremely short compared to the slow dynamics during phases 1 before and after the substitution, which appears to be instantaneous. With the parameter set in this plot it extends over a period of about only $3 \cdot 10^3$ generations. Parameters: $\nu_A = \nu_B = 30$; $s_1 = 0.0001$; $s_2 = 0.01$; $\mu = 10^{-6}$; $\epsilon = 10^{-6}$; $R = 0.0001$.

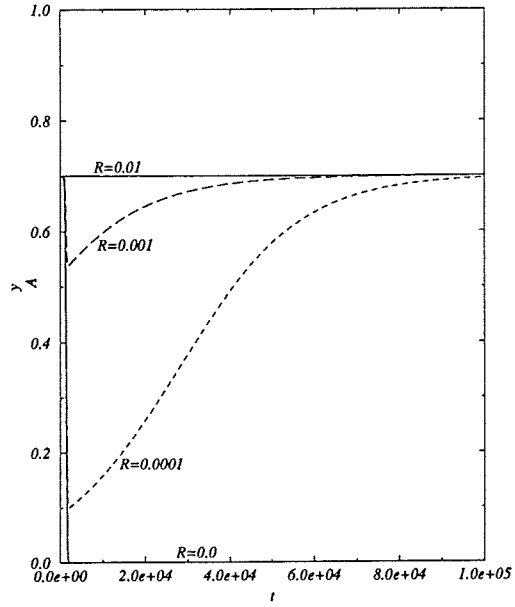

Figure 3.6: Recovery of master for different recombination rates. Directional selection model. Allele B occured together with a at time $t_1 = 0$. Parameters: $\nu_A = \nu_B = 30$; $s_1 = 10^{-4}$; $s_2 = 10^{-2}$; $\mu = 10^{-6}$; $R = 0, 10^{-4}, 10^{-3}, 10^{-2}$. If R is of the order of magnitude of s_2, the two loci evolve essentially independently (y_A does not drop below its equilibrium value).

Obviously, the two loci are independent and follow their 'own' mutation–selection dynamics if and only if $D = 0$, i.e. if they are in linkage equilibrium. As biological meaningful equilibria for phase 1 ($t < t_1$) we derive

$$\begin{aligned}(y_A, y_B, D)_1 &= (0, 0, 0) \text{ and} \\ (y_A, y_B, D)_2 &= (1 - \frac{\mu\nu_A}{s_1}, 0, 0)\,.\end{aligned}$$

The first equilibrium is attained only if replication accuracy is below the critical threshold value, which shall us not concern here. The two equilibria contain the threshold idea encountered in previous chapters in its simplest form. In cases where the back mutation rate may be neglected selection can only counteract the effects of mutation if the mutation rate is small enough ($\mu\nu_A < s_1$).

Now, for phase 2 we suppose that an a–b gamete mutates spontaneously at time $t = t_1$ to a–B. Thus, at time t_1 we have the following frequency distribution

$$y_A(t_1) = 1 - \frac{\mu\nu_A}{s_1}\,, \tag{3.15}$$

$$y_B(t_1) = \epsilon\,, \tag{3.16}$$

$$D(t_1) = \epsilon(1 - \frac{\mu\nu_A}{s_1})\,, \tag{3.17}$$

where $\epsilon << 1$ is the relative frequency of a–B gametes, just after being introduced into the population.

A complete analytical solution to system (3.12) to (3.17) is not known. We have to recede to approximations. First, we observe that in presence of selection at $\mathcal{B}$, selection at the linked $\mathcal{A}$ locus (via D) is negligible. Thus, we replace (3.13) by

$$\dot{y_B} = s_2 y_B(1 - y_B) - \mu\nu_B y_B\,. \tag{3.18}$$

The same argument leads to neglect of the quantity $s_1(1 - 2y_A)$ in (3.14). Furthermore, we omit terms involving 'second order parameters', i.e. the products $s_1 \cdot R$ and $s_2 \cdot R$. Therefore, Eq. (3.14) turns into

$$\dot{D} = D\big(s_2(1 - 2y_B) - R - \mu\nu_A - \mu\nu_B\big)\,. \tag{3.19}$$

Subject to the initial condition $y_B(t_1) = \epsilon$ the solution of (3.18) is straightforwardly

$$y_B(t) = \frac{\epsilon(1 - \frac{\mu\nu_B}{s_2})}{\epsilon + (1 - \epsilon - \frac{\mu\nu_B}{s_2})e^{(\mu\nu_B - s_2)(t - t_1)}}\,, \quad (t_1 < t < t_2)\,. \tag{3.20}$$

Inserting (3.20) into (3.19) variables can be separated. For $t_1 < t < t_2$, the solution of D, subject to $D(t_1) = \epsilon(1 - \frac{\mu\nu_A}{s_1})$, is

$$D(t) = D(t_1)e^{(\mu\nu_B - \mu\nu_A - s_2 - R)(t - t_1)}\Big(\frac{y_B(t - t_1)}{\epsilon}\Big)^2\,. \tag{3.21}$$

Finally, we replace Eq.(3.12) by

$$\dot{y}_A = -s_2 D - \mu\nu_{\mathcal{A}} y_A \,. \tag{3.22}$$

This equation is inhomogeneous linear and the solution is readily found to be

$$y_A(t) = \frac{y_A(t_1) - s_2 \int_{t_1}^{t} e^{\mu\nu_{\mathcal{A}}\tau} D(\tau) d\tau}{e^{\mu\nu_{\mathcal{A}}(t-t_1)}} \,. \tag{3.23}$$

Eq. (3.23) can be slightly simplified by observing that $e^{\mu\nu_{\mathcal{A}}(t-t_1)} \approx 1$ for $t_1 < t < t_2$, such that we may write

$$y_A(t) = y_A(t_1) - s_2 \int_{t_1}^{t} D(\tau) d\tau \,. \tag{3.24}$$

We would like to know $y_A(t_2)$ at the end of phase 2 ($t = t_2$). It would supply the initial condition for the new mutation–selection dynamics at $\mathcal{A}$ after the selective phase (phase 2) at locus $\mathcal{B}$ is completed. In particular, we may then obtain the recovery time for master A from the one-locus dynamics

$$\dot{y_A} = s_1 y_A (1 - y_A) - \mu\nu_{\mathcal{A}} y_A \,, \tag{3.25}$$

while $y_B = 1 - \mu\nu_{\mathcal{B}}/s_2$ and $D = 0$ remain constant. With our assumptions only one equilibrium exists. Therefore, convergence is global and the notion of 'recovery' not ambiguous. On integrating (3.25), subject to the initial condition

$$y_A(t_2) = \max(1 - \frac{\mu\nu_{\mathcal{A}}}{s_1} - s_2 \int_{t_1}^{t_2} D(\tau) d\tau, 0) \,, \tag{3.26}$$

we derive

$$t - t_2 = \frac{\log y_A(\tau) - \log(\mu\nu_{\mathcal{A}} - s_1 + s_1 y_A(\tau))}{s_1 - \mu\nu_{\mathcal{A}}} \Bigg|_{\tau=t_2}^{\tau=t} \,. \tag{3.27}$$

Replacing $y_A(t)$ by $1 - \frac{\mu\nu_{\mathcal{A}}}{s_1} - \tilde{\epsilon}$, the net recovery time T_* (i.e. measured from t_2 on) for the master to re-attain its equilibrium minus $\tilde{\epsilon}$

$$T_* = \frac{1}{s_1 - \mu\nu_{\mathcal{A}}} \log \frac{(1 - \frac{\mu\nu_{\mathcal{A}}}{s_1} - \tilde{\epsilon})(1 - \frac{\mu\nu_{\mathcal{A}}}{s_1} - y_A(t_2))}{\tilde{\epsilon} y_A(t_2)} \tag{3.28}$$

emerges. If $y_A(t_2)$ is identical *zero* the recovery time becomes infinite. This is correct, if, due to lacking back-flow, no A master can be restored. Furthermore, as is clear from formula (3.28), s_1 and T_* are inversely related: Stronger selection shortens the time to reach the equilibrium.

A master allele is not only in danger of being lost due to a replication mechanism which is too inaccurate to ensure its survival, but is also prone to loss if it is linked too tightly to neighboring sites. The master can be retained in the population however, if it has an opportunity to evolve independently of its neighbors. In other words, recombination can be a means to keep advantageous alleles, already been found, in the population. They do not have to be rebuilt by back-mutation after a selective sweep at a neighboring site and thus evolution may be accelerated.

For illustration we compare in Figure 3.7 to Figure 3.10 numerical solutions and analytical approximations. For these plots we chose a nucleotide mutation rate ($\mu = 10^{-6}$), which is below

Parameters				Numerical				Analytical			
				B occurs with A		B occurs with a		B occurs with A		B occurs with a	
s_1	s_2	R	μ	T^*	$y_A(t_2)$	T_*	$y_A(t_2)$	T^*	$y_A(t_2)$	T_*	$y_A(t_2)$
10^{-4}	10^{-2}	0	10^{-6}	$1.42 \cdot 10^5$	0.948	infinite	0.000	$1.42 \cdot 10^5$	1.	infinite	0.000
10^{-4}	10^{-2}	10^{-4}	10^{-6}	$1.41 \cdot 10^5$	0.917	$1.87 \cdot 10^5$	0.101	$1.41 \cdot 10^5$	0.961	$1.87 \cdot 10^5$	0.090
10^{-4}	10^{-2}	10^{-3}	10^{-6}	$1.26 \cdot 10^5$	0.764	$1.44 \cdot 10^5$	0.538	$1.26 \cdot 10^5$	0.775	$1.44 \cdot 10^5$	0.524
10^{-4}	10^{-2}	10^{-2}	10^{-6}	0.00	0.700	0.00	0.700	0.00	0.700	0.00	0.700
10^{-4}	10^{-2}	0	$3 \cdot 10^{-6}$	$9.12 \cdot 10^5$	0.851	infinite	0.000	$9.11 \cdot 10^5$	1.0	infinite	0.000
10^{-4}	10^{-2}	10^{-4}	$3 \cdot 10^{-6}$	$9.09 \cdot 10^5$	0.762	$1.11 \cdot 10^6$	0.013	$9.09 \cdot 10^5$	0.884	$1.11 \cdot 10^6$	0.013
10^{-4}	10^{-2}	10^{-3}	$3 \cdot 10^{-6}$	$8.83 \cdot 10^5$	0.305	$8.12 \cdot 10^5$	0.075	$8.84 \cdot 10^5$	0.326	$8.12 \cdot 10^5$	0.075
10^{-4}	10^{-2}	10^{-2}	$3 \cdot 10^{-6}$	$3.67 \cdot 10^4$	0.100	0.00	0.100	0.00	0.100	0.00	0.100
10^{-3}	10^{-2}	0	10^{-5}	$1.42 \cdot 10^4$	0.767	infinite	0.000	$1.42 \cdot 10^4$	1.000	infinite	0.000
10^{-3}	10^{-2}	10^{-3}	10^{-5}	$1.25 \cdot 10^4$	0.719	$1.47 \cdot 10^4$	0.625	$1.26 \cdot 10^4$	0.775	$1.44 \cdot 10^4$	0.524
10^{-3}	10^{-2}	10^{-2}	10^{-5}	0.00	0.700	0.00	0.700	0.00	0.700	0.00	0.700
$5 \cdot 10^{-3}$	$5 \cdot 10^{-1}$	0	10^{-4}	$5.05 \cdot 10^3$	0.899	infinite	0.000	$5.04 \cdot 10^3$	1.000	infinite	0.000
$5 \cdot 10^{-3}$	$5 \cdot 10^{-1}$	$5 \cdot 10^{-3}$	10^{-4}	$5.01 \cdot 10^3$	0.810	$6.04 \cdot 10^3$	0.079	$5.01 \cdot 10^3$	0.923	$6.25 \cdot 10^3$	0.052
$5 \cdot 10^{-2}$	$5 \cdot 10^{-1}$	$5 \cdot 10^{-2}$	10^{-3}	$4.26 \cdot 10^2$	0.425	$4.33 \cdot 10^2$	0.375	$4.65 \cdot 10^2$	0.550	$4.75 \cdot 10^2$	0.300

Table 3.1: Recovery times and frequencies of A after a selective sweep at $\mathcal{B}$. No back-flow. $\nu_{\mathcal{A}} = 30$, $\nu_{\mathcal{B}} = 30$, $\epsilon = 10^{-6}$, $\tilde{\epsilon} = 10^{-5}$.

the error threshold for the applied parameter set. It yields an equilibrium frequency of A of about 70 percent. Numerical solutions are obtained by integrating the system of gametic frequencies (3.2) by a standard integration routine. We used a fifth order Runge Kutta routine with adaptive step size.

The case that the advantageous mutation arises on a gamete carrying the master allele A can be treated analogously. We only have to take care of the different initial conditions for the dynamics of phase 2 at time $t = t_1$. They are

$$y_A(t_1) = 1 - \frac{\mu\nu_{\mathcal{A}}}{s_1}, \tag{3.29}$$

$$y_B(t_1) = \epsilon, \tag{3.30}$$

$$D(t_1) = -\epsilon\frac{\mu\nu_{\mathcal{A}}}{s_1}, \tag{3.31}$$

For $t_1 < t < t_2$ the same equations as before ((3.18), (3.19), (3.22)) describe the dynamics. Then, with the initial condition at time $t = t_2$

$$y_A(t_2) = \min(1 - \frac{\mu\nu_{\mathcal{A}}}{s_1} - s_2\int_{t1}^{t2} D(\tau)d\tau, 1) \tag{3.32}$$

we again integrate the mutation–selection equation (3.25), and obtain as net time for the master frequency y_A to return to its equilibrium value plus $\tilde{\epsilon}$

$$T^* = \frac{1}{s_1 - \mu\nu_{\mathcal{A}}}\log\frac{(1 - \frac{\mu\nu_{\mathcal{A}}}{s_1} + \tilde{\epsilon})(y_A(t_2) + \frac{\mu\nu_{\mathcal{A}}}{s_1} - 1)}{\tilde{\epsilon}y_A(t_2)}. \tag{3.33}$$

We would like to derive an 'easy' analytical expression for the master frequency of A at time $t = t_2$. Assuming that both loci have the same size, i.e. putting $\nu_{\mathcal{A}} = \nu_{\mathcal{B}}$, the exponential function in Eq. (3.21) simplifies to

$$e^{-(R+s_2)(t-t_1)}. \tag{3.34}$$

In the above scenario we assumed selection at $\mathcal{B}$ to be strong, i.e. $\mu\nu_{\mathcal{B}} << s_2$. Therefore, y_B is well approximated by

$$y_B \approx \frac{\epsilon}{\epsilon + (1-\epsilon)e^{-s_2(t-t_1)}}. \tag{3.35}$$

We write y_B^2 by means of y_B and $\dot{y_B}$. Subsequent integration by parts allows to get rid of the square in (3.21). This yields

$$\int_{t_1}^{t_2} D(t)dt \approx \frac{D(t_1)}{\epsilon s_2}\Big(-R\int_{t_1}^{t_2}\frac{e^{-Rt}}{\epsilon e^{s_2 t} + 1 - \epsilon}dt - \frac{e^{-Rt}}{\epsilon e^{s_2 t} + 1 - \epsilon}\Big|_{t_1}^{t_2}\Big). \tag{3.36}$$

As will be justified at a later occasion (see Chapter 4), the integral on the right hand side of the last expression is nearly identical to

$$R\int_{t_1}^{t_2}\frac{e^{-Rt}}{\epsilon e^{s_2 t} + 1 - \epsilon}dt \approx 1 - \epsilon^{\frac{R}{s_2}}. \tag{3.37}$$

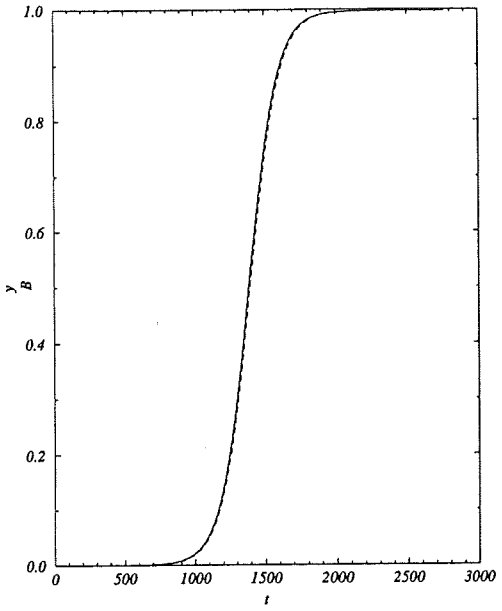

Figure 3.7: y_B during phase 2. Dotted: Numerical solution; B goes on A. Dashed: Numerical solution; B goes on a. Solid: Analytical approximation. Numerical solution and analytical approximation coincide very well independently of recombination rate. Parameters: $\nu_A = \nu_B = 30$; $\mu = 10^{-6}$; $s_1 = 10^{-4}$; $s_2 = 10^{-2}$; $R = 10^{-4}$.

We collect our approximate quantities, insert the initial condition $y_A(t_1)$ and derive

$$y_A(t_2) \approx y_A(t_1)\left(1 - \epsilon^{\frac{R}{s_2}} + \frac{\epsilon^{\frac{2R}{s_2}}}{1 + \frac{1}{\epsilon} - \epsilon}\right). \tag{3.38}$$

This can be simplified to

$$y_A(t_2) \approx y_A(t_1)\left(1 - \epsilon^{\frac{R}{s_2}}\right). \tag{3.39}$$

For the case that the advantageous mutation B occurs together with a master allele A one has to use the initial condition $D(t_1) = -\epsilon(1 - y_A(t_1))$. With analogous reasoning one derives

$$y_A(t_2) \approx y_A(t_1) + (1 - y_A(t_1))\epsilon^{\frac{R}{s_2}}. \tag{3.40}$$

Even though formulae (3.39) and (3.40) are only crude estimates, they make clear two things: Master A is the less affected by events happening at neighboring sites the higher the recombination rate is. In absence of recombination A gets wiped out. Secondly, a new advantageous mutation is possibly introduced in a single individual. Relating parameter ϵ to population size, it is natural to put $\epsilon = 1/2N$. Thus, the smaller the population the more important becomes recombination, since a minor reduction of the relative frequency of A may already mean extinction of A.

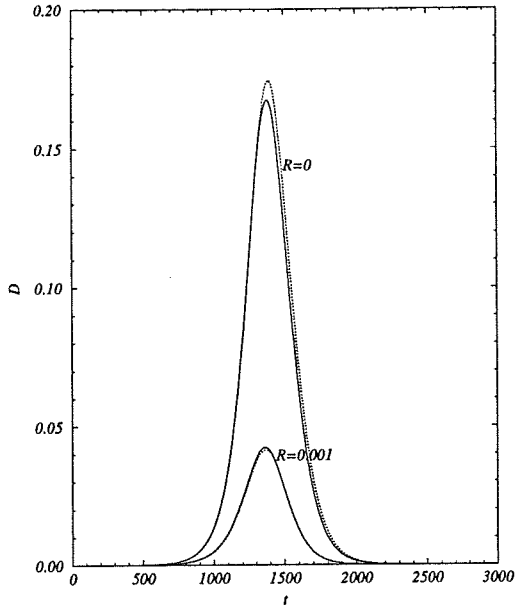

Figure 3.8: Linkage disequilibrium during phase 2. B goes on a. Dotted: Numerical solution. Solid: Analytical approximation. Parameters as before. Curves correspond to $R = 0.0$ and 10^{-3}, respectively. Before and after the selective phase 2 both loci are in linkage equilibrium or statistically independent.

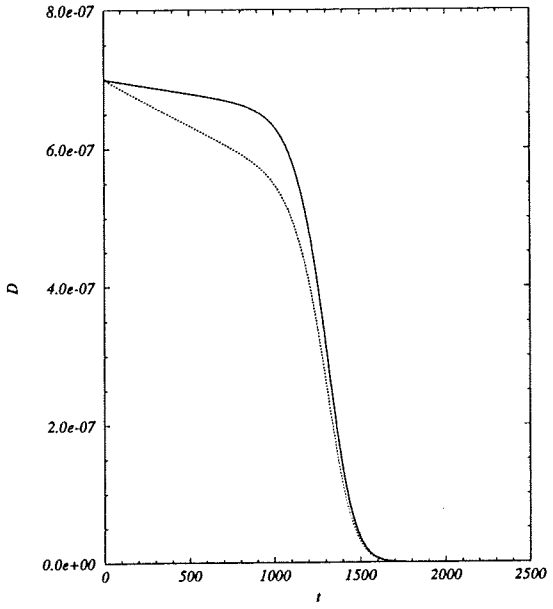

Figure 3.9: Linkage disequilibrium during phase 2. B goes on a. Dotted: Numerical solution. Solid: Analytical approximation. Parameters as before. $R = 0.01$.

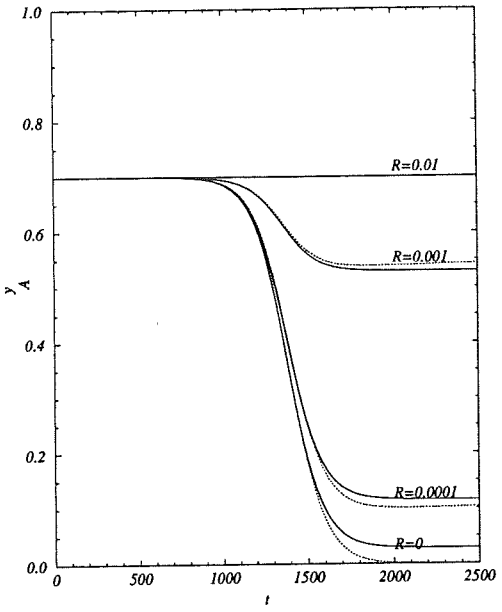

Figure 3.10: y_A during phase 2. B goes on a. Dotted: Numerical solution. Solid: Analytical approximation. Parameters as before. Curves correspond to $R = 10^{-2}$, 10^{-3}, 10^{-4}, $R = 0.0$.

3.4 The Recurrent Case

In the course of evolution advantageous mutations will be introduced into the population repeatedly. It is important to determine the long time effect of such events on population composition. We model advantageous mutations popping out recurrently at random times t_i, $i \in \mathbb{N}$, by a renewal process with parameter θ. In case that interarrival times between events are exponentially distributed, the renewal process can equivalently be described as Poisson counting process $(N_t)_t$, which counts the number of events happened until time t. This Poisson process also has rate θ, which means that each N_t follows a Poisson distribution with parameter θt. More precisely, we have

$$\text{Prob}(N_t = n) = e^{-\theta t}\frac{(\theta t)^n}{n!}. \tag{3.41}$$

We would like to find an answer to the question of how big the recombination fraction has at least to be in order not to loose master A when at neighboring sites substitutions occur repeatedly with rate θ. It is intuitively clear, that a high rate of evolution (i.e. rate of substitutions at neighboring sites) requires a high rate of recombination so that the master has the possibility of 'recombining away' from its neighboring sites which might drive it to extinction otherwise.

For the moment, we assume a 'worst case' scenario: Substitutions at $\mathcal{B}$ always form a–B gametes and thus lead to a reduction of the frequency of A. A more elaborate discussion, which takes also the possibility into account that occasional formation of A–B gametes can raise the frequency of A, will employ concepts from extreme value theory. We defer this to a later occasion.

Let $t = t_1 = 0$ denote the first time a substitution event takes place. It leads to a drop in master frequency such that $A_1 := y_A(t_1) = fA_0$, where A_0 is the mutation–selection equilibrium of A before t_1 and f is shorthand for $1 - \epsilon^{\frac{R}{s_2}}$. Let furthermore the sequence $(t_i)_{i>1}$ denote the (random-)times at which the subsequent substitution events happen and $(A_i)_{i>1}$ denote the according master frequencies 'right after' the substitutions took place. As noted before, the time span between introduction and establishment of strongly selected substitutions is very short. Compared to phase 1, phase 2 lasts only a time instant. To avoid notational difficulties we assume the counting process N_t to start recording events at time $t = \delta > 0$, so that we have $N_0 = 0$ – in accordance with standard notation. For times t such that $t_{i-1} \leq t < t_i$ the master dynamics is given by

$$y_A(t) = \frac{A_{i-1}(s_1 - \mu\nu_A)}{A_{i-1}s_1 + ((1 - A_{i-1})s_1 - \mu\nu_A)e^{(\mu\nu_A - s_1)(t-t_{i-1})}}. \tag{3.42}$$

First, we relate frequencies at two consecutive times t_{i-1} and t_i and ask under which condition the inequality

$$A_i \leq A_{i-1} \tag{3.43}$$

holds. If the inequality is satisfied then the event at time t_i means a deterioration of the situation with respect to the preceding event. Using the relation

$$A_i = y_A(t_i) = \frac{A_{i-1}(s_1 - \mu\nu_A)f}{A_{i-1}s_1 + ((1 - A_{i-1})s_1 - \mu\nu_A)e^{(\mu\nu_A - s_1)(t_i - t_{i-1})}}, \tag{3.44}$$

we find (3.43) to hold, iff

$$g(A_{i-1}) := \frac{1}{\mu\nu_A - s_1} \log\left(\frac{s_1(f - A_{i-1}) - \mu\nu_A f}{s_1(1 - A_{i-1}) - \mu\nu_A}\right) > \tau, \tag{3.45}$$

with τ defined as $t_i - t_{i-1}$. $g(x)$ is monotonically increasing in x with a singularity at $x = A_0 f$ and a minimum at $x = 0$ (see Figure 3.11). For this minimum one computes

$$g(0) := \hat{\tau} = \frac{1}{\mu\nu_A - s_1} \log f. \tag{3.46}$$

The interpretation of Figure 3.11 is the following. If waiting times between events $i - 1$ and i are longer than $\tau = g(A_{i-1})$ then the master has enough time to recover as to compensate for its drop in frequency due to the last selective event. If times are shorter however, then the master will enter a cascade of drops in frequency which finally drive it to extinction. The interarrival times of events are related to the rate θ of evolution by

$$\tau = \frac{1}{\theta}. \tag{3.47}$$

We note that a mean interarrival time of less than the critical value $\hat{\tau}$ will have devastating consequences for master A. In fact, as we will see in the sequel, a 0–1-law states that the mean interarrival time $\hat{\tau}$ separates survival from extinction.

As we know (see formula (3.42)), the frequency $y_A(t)$ at time t is a random variable (due to the random times at which substitution events occur). It can be written as a recursive function

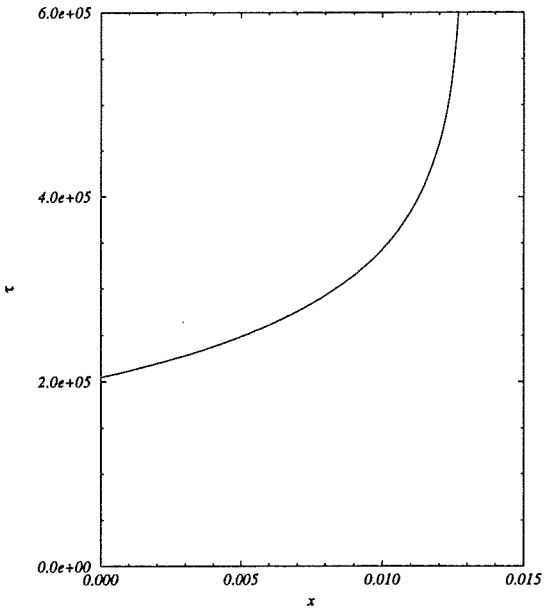

Figure 3.11: Function $g(x) = g(A_{i-1})$. Parameters: $\nu_{\mathcal{A}} = 30$; $s_1 = 10^{-4}$; $s_2 = 10^{-2}$; $\mu = 3 \cdot 10^{-6}$; $\epsilon = 10^{-6}$; $R = 10^{-4}$.

by means of the sequence $(A_i)_i$. A representation in closed form is not available, so we again have to confine ourselves to an approximation of the underlying process. Instead of modeling the $\mathcal{A}$–dynamics between events according to the logistic growth law (3.42), we approximate it by a simple exponential growth law, which is very accurate in the initial growth phase. With this assumption, one derives

$$A_i = f^i A_0 e^{(s_1 - \mu\nu_{\mathcal{A}})(t_i - t_1)} = A_0 f^{N_{t_i}+1} A_0 e^{(s_1-\mu\nu_{\mathcal{A}})t_i} \tag{3.48}$$

and for arbitrary times t

$$y_A(t) = A_0 f^{N_t+1} e^{(s_1-\mu\nu_{\mathcal{A}})t}\,. \tag{3.49}$$

We are interested in the long time ($t \to \infty$) fate of A. To that end we consider the probability that y_A does not fall below some positive value c ($0 < c < 1$). We have

$$\begin{aligned} \text{Prob}(y_A(t) \geq c) &= \text{Prob}(A_0 f \cdot f^{N_t} e^{(s_1-\mu\nu_{\mathcal{A}})t} \geq c) \\ &= \text{Prob}(N_t \leq \frac{\log \frac{c}{A_0 f} + (\mu\nu_{\mathcal{A}} - s_1)t}{\log f}) \\ &= \text{Prob}(N_t \leq \left[\frac{\log \frac{c}{A_0 f} + (\mu\nu_{\mathcal{A}} - s_1)t}{\log f}\right]) \\ &= \sum_{k=0}^{\left[\frac{\log \frac{c}{A_0 f}}{\log f} + \hat{\theta} t\right]} e^{-\theta t} \frac{(\theta t)^k}{k!} \end{aligned} \tag{3.50}$$

$$= \sum_{k=0}^{\left[\frac{\log \frac{c}{A_0 f}}{\log f}+\tau\right]} e^{-\frac{\theta}{\hat{\theta}}\tau} \frac{(\frac{\theta}{\hat{\theta}}\tau)^k}{k!},$$

where we used the abbreviation $\hat{\theta} := \hat{\tau}^{-1} = \frac{\mu\nu_A - s_1}{\log f}$ and, in the last equality, substituted $\hat{\theta}t = \tau$. Brackets [.], as usual, denote the largest integer smaller than the argument inside the brackets.

The first summand (being a small constant) in the upper summation bound in the last two expressions does not affect in any way limiting behavior of the sum when passing to large t. That means, what we in fact would like to know is the value of

$$\lim_{\tau\to\infty} e^{-\frac{\theta}{\hat{\theta}}\tau} \sum_{k=0}^{\tau} \frac{(\frac{\theta}{\hat{\theta}}\tau)^k}{k!}. \tag{3.51}$$

Two things are remarkable about this last expression. First, it turns out that the limit obeys a 0–1 law with the barrier set by the fraction $\frac{\theta}{\hat{\theta}}$. More precisely,

$$\lim_{\tau\to\infty} e^{-\frac{\theta}{\hat{\theta}}\tau} \sum_{k=0}^{\tau} \frac{(\frac{\theta}{\hat{\theta}}\tau)^k}{k!} = \begin{cases} 0, & \text{if } \theta > \hat{\theta}, \\ 1, & \text{if } \theta < \hat{\theta}. \end{cases} \tag{3.52}$$

We show derivation of Eq. (3.52) in the Appendix.

Secondly, the above behavior does not explicitly depend on c. That means, we can infer that in the long time limit the master frequency will be below any positive value c with probability *one* if the rate fulfills $\theta > \hat{\theta}$. This is equivalent to the statement that the master will become extinct with probability *one* in this case ($\text{Prob}(y_A(t = \infty) = 0) = 1$). On the other hand, if selective events happen at a rate $\theta < \hat{\theta}$, the master will manage to survive: $\text{Prob}(y_A(t = \infty) = 0) = 0$. Due to the assumptions of our model we cannot make this statement more precise, since the exponential growth property obscures the fact that relative frequencies have to be restricted to at most the interval $[0, 1]$.

Having recognized the important role of the threshold $\hat{\theta}$, we use this expression to derive a minimal recombination rate, which is necessary for master A to survive recurrent substitutions in its neighborhood. We only have to solve

$$\hat{\theta} = \frac{\mu\nu_A - s_1}{\log f} \tag{3.53}$$

for recombination rate R and find

$$R_{min} = \frac{s_2 \log(1 - e^{\frac{\mu\nu_A - s_1}{\theta}})}{\log \epsilon}. \tag{3.54}$$

We noted already that ϵ may be related to population size via $\epsilon = \frac{1}{2N}$. Then Eq. (3.54) shows that small populations require a higher minimal recombination rate – a result which, as one expects, carries over from the non-recurrent case treated before. Furthermore, stronger selection at neighboring sites calls for more chances for recombination: The dependence of R_{min} on s_2

is positively linear. On the other hand, also intuitively clear, increasing selection for A lowers minimal required recombination rate. To first order, R_{min} decreases in s_1 with a coefficient of the order $O(-2/\theta)$.

One may ask what happens if $O(s_1) = O(s_2)$. We do not elaborate on this case here, but only remark that – under directional selection at both loci – recombination facilitates coexistence of the two advantageous alleles A and B in the population.

3.5 Simulation Results

To get an idea how finite population size affects the above analytical result for infinitely large populations, we carried out Monte Carlo simulations using a Wright-Fisher model (Ewens, 1979). This amounts to multinomial sampling of any new generation from the previous one, where population size is kept fixed from generation to generation. In Figure 3.12 we plotted extinction probability of master A versus recombination rate R. For each choice of R we averaged over 100 trajectories and recorded how many times master A got lost within the course of $2 \cdot 10^4$ generations. As expected, the discontinuous 0–1-law is smoothened in dependence of population size. For recombination rates above R_{min} one still has a positive probability of extinction, since any allele, present in low concentration, is prone to loss by random drift, which is neglected in the analytical formula Eq. (3.54). The effect of random drift is the stronger the smaller the population is.

We remark, that the simulations take both possibilities into account; strongly selected alleles B may arise on a gamete carrying allele A *or* allele a. The probability of realizing the one or the other is given by the actual relative frequency of a or A, respectively.

3.6 Summary and Extensions

Let's finally be clear again on what we have neglected and what cannot be inferred from relation (3.54). We did not take into account the possibility that B alleles may occur together with A and thus form an A–B gamete. This would disrupt the cascade of lowering the frequencies of A. Thus, R_{min} above overestimates the actual minimal recombination rate. However, as stated earlier, we are interested to give an account of the 'worst case' scenario (which implies assuming a series of formation of a–B gametes).

Another point is that we always assumed recombination rate to be constant for all and ever. This is definitely not realistic. Rather, one has to think of the series of substitutions to affect the neighborhood of $\mathcal{A}$. That means any two substitutions will occur at different sites near $\mathcal{A}$ and not at the same (recombinational) distance. This amounts to modeling recombination rate as a stochastic quantity or at least to take some form of average (cf. Chapter 4), reflecting the different distances of '$\mathcal{B}$-loci' from $\mathcal{A}$. So, what one should think of when talking about the recombination

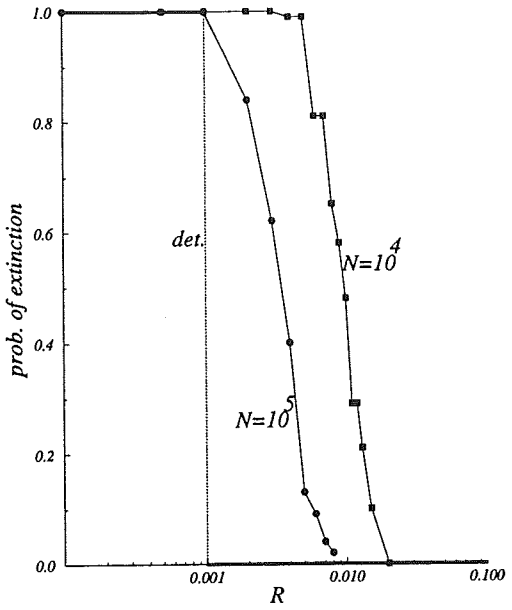

Figure 3.12: Probability of extinction of master A with recurring substitutions at a neighboring locus $\mathcal{B}$. Dotted: Analytical value according to Eq. (3.54). Solid: Simulation results based on 100 repeats over $2 \cdot 10^4$ generations each. Initial population consisted of A–b gametes only. Parameters: $\nu_{\mathcal{A}} = 30$, $\nu_{\mathcal{B}} = 30$, $s_1 = 0.05$, $s_2 = 0.5$, $\mu = 0.001$, $\theta = 0.005$.

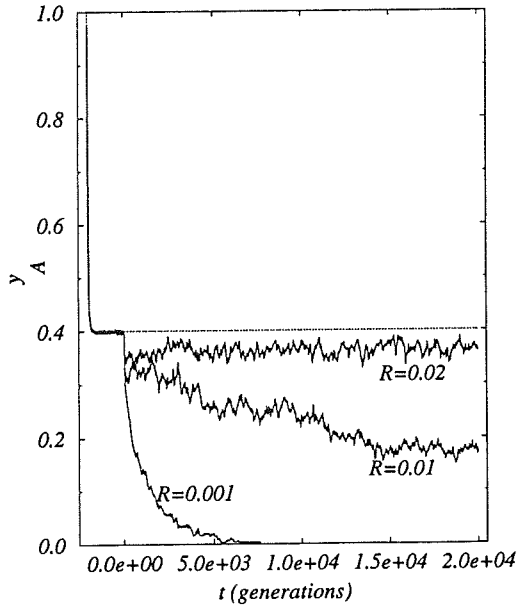

Figure 3.13: Frequency of master A decreases more quickly for low recombination rates. Simulation results averaged over 100 trajectories. Simulations were started at time $t = -2 \cdot 10^3$ with a homogeneous population consisting of gametes A-b only. Up to time $t = 0$ a-b gametes accumulate through mutation at locus $\mathcal{A}$ until mutation-selection balance (here at a frequency of A at about $y_A = 0.4$) is reached. After $t = 0$ strongly selected mutants B are introduced according to a Poisson process with rate θ. Parameters: $\nu_{\mathcal{A}} = 30$, $\nu_{\mathcal{B}} = 30$, $s_1 = 0.05$, $s_2 = 0.5$, $\mu = 0.001$, $\theta = 0.005$.

rate in the above exposition, is a *mean* of a stochastic quantity rather than a fixed deterministic rate.

Our considerations show that recombination may be disadvantageous or advantageous, depending on the particular fitness landscape. Where functionality depends on multiple gene sites, such as in coadapted gene complexes (Wright, 1955), recombination may distort favorable combinations. One would expect recombination rates to be reduced in such regions. The adequate model would be the one introduced in section 3.2.

On the other hand, functionally independent genes may profit from the possibility of recombination. Favorable constellations at individual sites are more easily retained or established, if evolution at single sites is independent of the chromosomal neighborhood. This is the more important the higher the rate of evolution is. Such independence can be gained by recombination – the adequate model is the one of sections 3.3 and 3.4.

To our knowledge, as to the present time there are no data available which relate functionality of genes or gene complexes and recombination rate. Due to the present lack of experimental data corroboration of these hypotheses is an open point.

4

Effect of Strong Selection

We study again a two locus - two allele model. The point of view will be different: Our interest is focussed on the effect a selective substitution at a second gene locus has on *polymorphism* at a first site. Both sites, as we assume, are linked more or less tightly depending on the recombinational distance between them and the selective advantage of the substituting allele.

4.1 The Deterministic Approach

Maynard Smith & Haigh (1974) were the first to study this so–called 'Hitchhiking Effect' in a deterministic (infinite population sizes) setting.

It turns out to be convenient for the following to transform Eq.s (3.2) into a system with equations for $y_B = x_1 + x_2$, $z_{A|B}$ and $z_{A|b}$. The latter two quantities are conditional relative frequencies. $z_{A|B}$ ($z_{A|b}$) describes the fraction of A alleles given the condition they carry B (b) at locus $\mathcal{B}$. We assume the following scenario: A neutral polymorphism of two alleles, A and a, is present at locus $\mathcal{A}$. I.e. there is no selective difference between individuals carrying one or the other allele. At locus $\mathcal{B}$ a strongly advantageous allele B will be introduced into the population at time $t_1 = 0$. Before t_1 only b alleles are present. Besides these occasional mutation events leading to some advantageous B, no other mutation forces are acting. This model is an extreme case of the one in the previous chapter. Mutation at $\mathcal{B}$ is outweighed by strong selection and therefore negligible, mutation at $\mathcal{A}$ is irrelevant for what concerns fitness of an individual due to neutrality. The fitness matrix reads

$$\tilde{W} = \begin{pmatrix} 1+2s_2 & 1+2s_2 & 1+s_2 & 1+s_2 \\ 1+2s_2 & 1+2s_2 & 1+s_2 & 1+s_2 \\ 1+s_2 & 1+s_2 & 1 & 1 \\ 1+s_2 & 1+s_2 & 1 & 1 \end{pmatrix}. \tag{4.1}$$

Thus, the only driving forces of evolution are assumed to be strong selection at $\mathcal{B}$ and random drift at $\mathcal{A}$ (the latter acts only in finite populations, see section 2). For this reason the dynamics at $\mathcal{B}$ is modeled as deterministic, determined by the ordinary selection equation. The recombination rate R describes linkage between the two loci. With these assumptions model (3.2) transforms equivalently to

$$\dot{z}_{A|B} = R(1+s_2)(1-y_B)(z_{A|b}-z_{A|B}), \tag{4.2}$$

$$\dot{z}_{A|b} = R(1+s_2)y_B(z_{A|B}-z_{A|b}), \tag{4.3}$$

$$\dot{y}_B = s_2 y_B(1-y_B). \tag{4.4}$$

Clearly, we again approximate the product $R(1+s_2)$ by R. The solution of (4.4), subject to the initial condition $y_B(t_1)=\epsilon$ is

$$y_B(t) = \frac{\epsilon}{\epsilon+(1-\epsilon)e^{-s_2(t-t_1)}}. \tag{4.5}$$

We would like to know by which amount *heterozygosity* H at $\mathcal{A}$ is affected when at $\mathcal{B}$ an allele B is introduced and subsequently fixed (a so–called *substitution*). To this end we need to know $H(t) := 2y_A(t)(1-y_A(t))$ at times $t=t_1$ and t_2. The latter is the time when the selection process at $\mathcal{B}$ is completed, i.e. $y_B(t_2)=1-\epsilon$. Solutions to equations (4.2) and (4.3) can be obtained in the integral form

$$z_{A|B}(t) = z_{A|B}(t_1) - R(z_{A|B}(t_1)-z_{A|b}(t_1))\int_{t_1}^{t}\frac{(1-\epsilon)e^{-(s_2+R)(\tau-t_1)}}{\epsilon+(1-\epsilon)e^{-s_2(\tau-t_1)}}d\tau, \tag{4.6}$$

$$z_{A|b}(t) = z_{A|b}(t_1) + R(z_{A|B}(t_1)-z_{A|b}(t_1))\int_{t_1}^{t}\frac{\epsilon e^{-R(\tau-t_1)}}{\epsilon+(1-\epsilon)e^{-s_2(\tau-t_1)}}d\tau. \tag{4.7}$$

The initial conditions have to be chosen according to whether the advantageous mutant B arises at an A– or an a–carrying allele. In the first case we have

$$z_{A|B}(t_1)=1 \text{ and } z_{A|b}(t_1)=\frac{y_A(t_1)-\epsilon}{1-\epsilon}\approx y_A(t_1) \tag{4.8}$$

and in the second case

$$z_{A|B}(t_1)=0 \text{ and } z_{A|b}(t_1)=\frac{y_A(t_1)}{1-\epsilon}\approx y_A(t_1). \tag{4.9}$$

Of course, y_A means again frequency of A alleles. The approximation holds, if ϵ is small compared to $y_A(t_1)$.

When computing the magnitude of the hitchhiking effect we have to account for both possibilities, B uniting either with A or with a. We do this by calculating a weighted average

$$\bar{H}(t) := y_A(t_1)H(t)|_{z_{A|B}(t_1)=1} + (1-y_A(t_1))H(t)|_{z_{A|B}(t_1)=0}. \tag{4.10}$$

In terms of the new quantities y_A is computed as

$$y_A(t) = z_{A|B}(t)y_B(t) + z_{A|b}(t)(1-y_B(t)). \tag{4.11}$$

Therefore, at time $t = t_2$ we have

$$y_A(t_2) = z_{A|B}(t_2) + \epsilon(z_{A|b}(t_2) - z_{A|B}(t_2)), \tag{4.12}$$

since $y_B(t_2) = 1 - \epsilon$. We apply relations (4.6) and (4.7) and derive

$$\begin{aligned} & y_A(t_2) = \\ & z_{A|B}(t_2) + \epsilon(z_{A|b}(t_1) - z_{A|B}(t_1))\Big(1 - R\int_{t_1}^{t_2} \frac{e^{-(s_2+R)(\tau-t_1)}}{\epsilon + (1-\epsilon)e^{-s_2(\tau-t_1)}}(1 - \epsilon + \epsilon e^{s_2(\tau-t_1)})d\tau\Big) = \\ & z_{A|B}(t_2) + \epsilon(z_{A|b}(t_1) - z_{A|B}(t_1))\Big(1 - R\int_{t_1}^{t_2} e^{-R(\tau-t_1)}d\tau\Big) = \\ & z_{A|B}(t_2) + (z_{A|b}(t_1) - z_{A|B}(t_1))\epsilon^{1+\frac{2R}{s_2}}. \end{aligned} \tag{4.13}$$

In the last equation we used that

$$t_2 = -\frac{2}{s_2}\log(\epsilon), \tag{4.14}$$

which is obtained from Eq. (4.5). We now express reduction of heterozygosity due to the substitution event as

$$\frac{\bar{H}(t_2)}{\bar{H}(t_1)} = \frac{\bar{H}(t_2)}{2z_{A|b}(t_1)(1 - z_{A|b}(t_1))}. \tag{4.15}$$

Inserting $y_A(t_2)$ according to (4.13) into the formula for H, and this in turn into (4.10), we obtain after some calculations

$$\frac{\bar{H}(t_2)}{\bar{H}(t_1)} = R \cdot I(t_2)(2 - R \cdot I(t_2)) + \epsilon^{1+\frac{2R}{s_2}}(2 - \epsilon^{1+\frac{2R}{s_2}}) - 2R \cdot \epsilon^{1+\frac{2R}{s_2}} I(t_2), \tag{4.16}$$

with

$$I(t_2) := \int_{t_1}^{t_2} \frac{(1-\epsilon)e^{-(s_2+R)(\tau-t_1)}}{\epsilon + (1-\epsilon)e^{-s_2(\tau-t_1)}}d\tau. \tag{4.17}$$

The integral in the last equation can be very well approximated. We note that Eq. (4.17) describes essentially the time dependent behavior of $z_{A|B}$. This quantity changes significantly only for $t < \frac{t_2}{2}$. In this time domain the denominator in (4.17) may be replaced by $e^{-s_2(\tau-t_1)}$. Therefore, we approximate

$$I(t_2) \approx I(\frac{t_2}{2}) \approx \int_{t_1}^{\frac{t_2}{2}} e^{-R(\tau-t_1)}d\tau = \frac{1}{R}(1 - \epsilon^{\frac{R}{s_2}}). \tag{4.18}$$

By the way, these remarks provide the still open justification for the same approximation which had been used in the previous chapter.

We insert the latter result into (4.16) to get

$$\begin{aligned} & \frac{\bar{H}(t_2)}{\bar{H}(t_1)} = \\ & (1 - \epsilon^{\frac{R}{s_2}})(1 + \epsilon^{\frac{R}{s_2}}) + \epsilon^{1+\frac{3R}{s_2}}(2 - \epsilon^{1+\frac{R}{s_2}}) = \\ & 1 - \epsilon^{\frac{2R}{s_2}}(1 + \epsilon^{1+\frac{R}{s_2}}(2 - \epsilon^{1+\frac{R}{s_2}})) = \\ & 1 - \epsilon^{\frac{2R}{s_2}} + O(\epsilon^{1+\frac{3R}{s_2}}). \end{aligned} \tag{4.19}$$

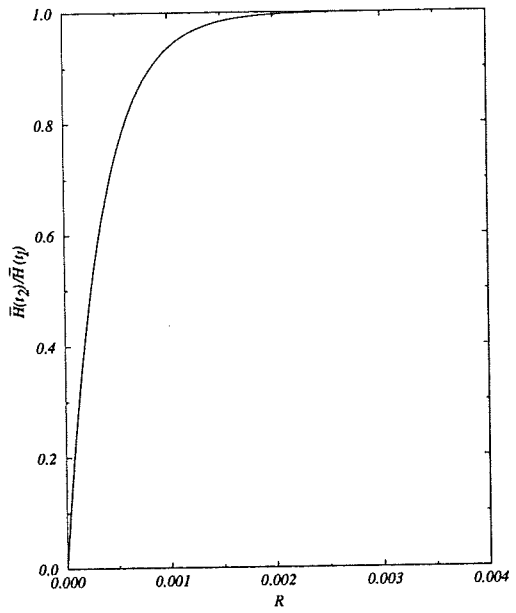

Figure 4.1: Reduction of heterozygosity at $\mathcal{A}$ due to a single substitution at $\mathcal{B}$ in dependence of recombination rate R according to formula (4.19). Parameters: $s_2 = 10^{-2}$; $\epsilon = 10^{-6}$.

In Figure 4.1 we show how the hitchhiking effect, as given by reduction in heterozygosity via formula (4.19), depends on recombination rate. As is clear from Eq. (4.19), a recombination rate bigger than half of the selective value s_2 means that the allelic composition of the $\mathcal{A}$ locus is barely affected by a substitution at $\mathcal{B}$. This property is fairly robust with respect to ϵ. As a rule of thumb we may repeat that evolution of two adjacent gene loci proceeds independently, if recombinational and selective forces, as expressed by their according parameters, are of the same order of magnitude.

4.2 Finite Population Size

We model the effect of random genetic drift due to sampling fluctuations from one generation to the next by a diffusion process on a suitable state space. The diffusion process is characterized by a differential operator $\mathcal{L}$. We give this in its general form for a two-locus two-allele model, which includes mutation, selection and recombination. The original formulations for gametic frequencies are due to Ohta & Kimura (1969) and, more generally, Ethier & Nagylaki (1989). We recall the notation from Chapter 3 (p.49) for the four gametic types. Since the relative frequency of one of them is determined by the remaining three, we deal with a diffusion process on state space $[0,1]^3$. The differential operator $\mathcal{L}$ and its adjoint operator $\mathcal{L}^*$ constitute the right hand sides of

the Kolmogorov forward and backward equations. $\mathcal{L}$ is given by

$$\mathcal{L} = \frac{1}{2}\sum_{i,j=1}^{3} a_{ij}(x)\frac{\partial^2}{\partial x_i \partial x_j} + \sum_{i,j=1}^{3} b_i(x)\frac{\partial}{\partial x_i}, \tag{4.20}$$

with diffusion and drift coefficients $a_{ij}(x)$ and $b_i(x)$, respectively. More explicitly, including forward and backward mutation, (directional) selection and recombination, the forward operator equation reads

$$\begin{aligned}
&\frac{\partial}{\partial t} = \qquad (4.21)\\
&\Big(- x_1(\mu_{\mathcal{A}} + \mu_{\mathcal{B}}) + x_2\bar{\mu}_{\mathcal{A}} + \\
&\quad s_1 x_1(1 - x_1 - x_3) + s_2 x_1(1 - x_1 - x_2) + Rw_{14}D\Big)\frac{\partial}{\partial x_1} + \\
&\Big(- x_2(\bar{\mu}_{\mathcal{A}} + \mu_{\mathcal{B}}) + x_1\mu_{\mathcal{B}} + x_4\bar{\mu}_{\mathcal{B}} - \\
&\quad s_1 x_2(x_1 + x_3) + s_2 x_2(1 - x_1 - x_2) - Rw_{14}D\Big)\frac{\partial}{\partial x_2} + \\
&\Big(- x_3(\mu_{\mathcal{A}} + \bar{\mu}_{\mathcal{B}}) + x_4\bar{\mu}_{\mathcal{A}} + x_1\mu_{\mathcal{B}} + \\
&\quad s_1 x_3(1 - x_1 - x_3) - s_2 x_3(x_1 + x_2) - Rw_{14}D\Big)\frac{\partial}{\partial x_3} + \\
&\frac{1}{4N}\Big(x_1(1 - x_1)\frac{\partial^2}{\partial x_1^2} - 2x_1x_2\frac{\partial^2}{\partial x_1 \partial x_2} - 2x_1x_3\frac{\partial^2}{\partial x_1 \partial x_3} - 2x_2x_3\frac{\partial^2}{\partial x_2 \partial x_3} + \\
&\quad x_2(1 - x_2)\frac{\partial^2}{\partial x_2^2} + x_3(1 - x_3)\frac{\partial^2}{\partial x_3^2}\Big).
\end{aligned}$$

Here, $\mu_{\mathcal{A}}$ and $\mu_{\mathcal{B}}$ are forward mutation rates and $\bar{\mu}_{\mathcal{A}}$ and $\bar{\mu}_{\mathcal{B}}$ backward rates (i.e mutation from a to A and b to B, respectively). The former may be identified with $\mu\nu_{\mathcal{A}}$ and $\mu\nu_{\mathcal{B}}$ from the previous chapter where the size of the two loci was explicitly taken into account.

For the hitchhiking model a representation of $\mathcal{L}$ in terms of y_B, $z_{A|B}$ and $z_{A|b}$ is desirable. We transform the above partial derivatives into partial derivatives with respect to the new variables $z_{A|B}$, $z_{A|b}$, y_B. Applying the chain rule we derive

— first order differentials —

$$\frac{\partial}{\partial x_1} = \frac{1 - z_{A|B}}{y_B}\frac{\partial}{\partial z_{A|B}} + \frac{z_{A|b}}{1 - y_B}\frac{\partial}{\partial z_{A|b}} + \frac{\partial}{\partial y_B} \tag{4.22}$$

$$\frac{\partial}{\partial x_2} = \frac{-z_{A|B}}{y_B}\frac{\partial}{\partial z_{A|B}} + \frac{z_{A|b}}{1 - y_B}\frac{\partial}{\partial z_{A|b}} + \frac{\partial}{\partial y_B} \tag{4.23}$$

$$\frac{\partial}{\partial x_3} = \frac{1}{1 - y_B}\frac{\partial}{\partial z_{A|b}} \tag{4.24}$$

— second order differentials —

$$\begin{aligned}
\frac{\partial^2}{\partial x_1^2} = &\frac{-2(1 - z_{A|B})}{y_B^2}\frac{\partial}{\partial z_{A|B}} + \frac{2z_{A|b}}{(1 - y_B)^2}\frac{\partial}{\partial z_{A|b}} + \\
&(\frac{1 - z_{A|B}}{y_B})^2\frac{\partial^2}{\partial z_{A|B}^2} + (\frac{z_{A|b}}{1 - y_B})^2\frac{\partial^2}{\partial z_{A|b}^2} + \frac{\partial^2}{\partial y_B^2} +
\end{aligned}$$

$$\frac{2z_{A|b}(1-z_{A|B})}{y_B(1-y_B)}\frac{\partial^2}{\partial z_{A|B}\partial z_{A|b}}+\frac{2(1-z_{A|B})}{y_B}\frac{\partial^2}{\partial z_{A|B}\partial y_B}+\frac{2z_{A|b}}{1-y_B}\frac{\partial^2}{\partial z_{A|b}\partial y_B} \quad (4.25)$$

$$\frac{\partial^2}{\partial x_2^2} = \frac{2z_{A|B}}{y_B^2}\frac{\partial}{\partial z_{A|B}}+\frac{2z_{A|b}}{(1-y_B)^2}\frac{\partial}{\partial z_{A|b}}+(\frac{z_{A|B}}{y_B})^2\frac{\partial^2}{\partial z_{A|B}^2}+(\frac{z_{A|b}}{1-y_B})^2\frac{\partial^2}{\partial z_{A|b}^2}+\frac{\partial^2}{\partial y_B^2}+$$
$$(\frac{-2z_{A|B}z_{A|b}}{y_B(1-y_B)}\frac{\partial^2}{\partial z_{A|B}\partial z_{A|b}}+\frac{-2z_{A|B}}{y_B}\frac{\partial^2}{\partial z_{A|B}\partial y_B}+\frac{2z_{A|b}}{1-y_B}\frac{\partial^2}{\partial z_{A|b}\partial y_B} \quad (4.26)$$

$$\frac{\partial^2}{\partial x_3^2} = (\frac{1}{1-y_B})^2\frac{\partial^2}{\partial z_{A|b}^2} \quad (4.27)$$

— mixed second order differentials —

$$\frac{\partial^2}{\partial x_1\partial x_2} = \frac{2z_{A|B}-1}{y_B^2}\frac{\partial}{\partial z_{A|B}}+\frac{2z_{A|b}}{(1-y_B)^2}\frac{\partial}{\partial z_{A|b}}+$$
$$\frac{-(1-z_{A|B})z_{A|B}}{y_B^2}\frac{\partial^2}{\partial z_{A|B}^2}+(\frac{z_{A|b}}{1-y_B})^2\frac{\partial^2}{\partial z_{A|b}^2}+\frac{\partial^2}{\partial y_B^2}+$$
$$\frac{z_{A|b}(1-2z_{A|B})}{y_B(1-y_B)}\frac{\partial^2}{\partial z_{A|B}\partial z_{A|b}}+\frac{1-2z_{A|B}}{y_B}\frac{\partial^2}{\partial z_{A|B}\partial y_B}+\frac{2z_{A|b}}{1-y_B}\frac{\partial^2}{\partial z_{A|b}\partial y_B} \quad (4.28)$$

$$\frac{\partial^2}{\partial x_1\partial x_3} = (\frac{1}{1-y_B})^2\frac{\partial}{\partial z_{A|b}}+\frac{z_{A|b}}{(1-y_B)^2}\frac{\partial^2}{\partial z_{A|b}^2}+$$
$$\frac{1-z_{A|B}}{y_B(1-y_B)}\frac{\partial^2}{\partial z_{A|B}\partial z_{A|b}}+\frac{1}{1-y_B}\frac{\partial^2}{\partial z_{A|b}\partial y_B} \quad (4.29)$$

$$\frac{\partial^2}{\partial x_2\partial x_3} = (\frac{1}{1-y_B})^2\frac{\partial}{\partial z_{A|b}}+\frac{z_{A|b}}{(1-y_B)^2}\frac{\partial^2}{\partial z_{A|b}^2}+$$
$$\frac{-z_{A|B}}{y_B(1-y_B)}\frac{\partial^2}{\partial z_{A|B}\partial z_{A|b}}+\frac{1}{1-y_B}\frac{\partial^2}{\partial z_{A|b}\partial y_B}. \quad (4.30)$$

Inserting these expressions into (4.20), we obtain for the differential operator $\mathcal{L}$ in its new coordinates

$$\mathcal{L} = \frac{1}{4N}\left(\frac{z_{A|B}(1-z_{A|B})}{y_B}\frac{\partial^2}{\partial z_{A|B}^2}+\frac{z_{A|b}(1-z_{A|b})}{1-y_B}\frac{\partial}{\partial z_{A|b}^2}+y_B(1-y_B)\frac{\partial}{\partial y_B^2}\right)+$$
$$\Big(R(1-y_B)(z_{A|b}-z_{A|B})+s_1z_{A|B}(1-z_{A|B})-$$
$$\mu_{\mathcal{A}}z_{A|B}+\bar{\mu}_{\mathcal{A}}(1-z_{A|B})+\bar{\mu}_{\mathcal{B}}(z_{A|b}-z_{A|B})\frac{1-y_B}{y_B}\Big)\frac{\partial}{\partial z_{A|B}}+$$
$$\Big(Ry_B(z_{A|B}-z_{A|b})+s_1z_{A|b}(1-z_{A|b})$$
$$-\mu_{\mathcal{A}}z_{A|b}+\bar{\mu}_{\mathcal{A}}(1-z_{A|b})+\mu_{\mathcal{B}}(z_{A|B}-z_{A|b})\frac{y_B}{1-y_B}\Big)\frac{\partial}{\partial z_{A|b}}+$$
$$\Big(y_B(1-y_B)(s_2+s_1(z_{A|B}-z_{A|b}))-\mu_{\mathcal{B}}y_B+\bar{\mu}_{\mathcal{B}}(1-y_B)\Big)\frac{\partial}{\partial y_B}. \quad (4.31)$$

Now, we specialize to the hitchhiking scenario introduced before. Neutrality at $\mathcal{A}$, strong selection at $\mathcal{B}$, which outweighs also random drift. We therefore treat the dynamics at $\mathcal{B}$ deterministically. This is easy to accomplish, since the transformation above provides a separate equation for B. So we only have to omit the diffusion term involving the y_B derivative in the above operator.

Furthermore, neglect of mutation and neutrality at $\mathcal{A}$ leads to $s_1 = \mu_{\mathcal{A}} = \mu_{\mathcal{B}} = \bar{\mu}_{\mathcal{A}} = \bar{\mu}_{\mathcal{B}} = 0$. The forward equation is satisfied by the transition density $\phi(z^{(0)}_{A|B}, z^{(0)}_{A|b}, y^{(0)}_B, z_{A|B}, z_{A|b}, y_B, t)$. The latter measures the probability to go in time span t from position $(z^{(0)}_{A|B}, z^{(0)}_{A|b}, y^{(0)}_B)$ to a position $(z_{A|B}, z_{A|b}, y_B) \in S$ in state space, where $S \subset [0,1]^3$, i.e.

$$\text{Prob}\big((Z_{A|B}(t), Z_{A|b}(t), y_B(t)) \in S \mid z^{(0)}_{A|B}, z^{(0)}_{A|b}, y^{(0)}_B\big) = \int_S d\phi(z^{(0)}_{A|B}, z^{(0)}_{A|b}, y^{(0)}_B, z_{A|B}, z_{A|b}, y, t)\,. \tag{4.32}$$

As well, the forward equation may be integrated on both sides over some function $f = f(z_{A|B}, z_{A|b}, y_B)$ with respect to this density; the equation still holds – thereby changing from a partial to an ordinary differential equation (see Appendix). Since Z_i are random variables now, they are denoted by capital letters, as opposed to the deterministic quantities z_i. We apply the differential operator (4.31) to the special functions $f = z_{A|B}$, $f = z_{A|b}$ and $f = y_B$ and choose $S = [0,1]^3$. Thus, the integral over f with respect to ϕ yields expectations of the random variables $Z_{A|B}$ and $Z_{A|b}$. y_B is, according to our assumption, no random variable, therefore $E(Y_B) = y_B$. Integrating $\frac{\partial f}{\partial t} = \mathcal{L}f$ with the above choices for f gives the results

$$\frac{dE(Z_{A|B})}{dt} = R(1-y_B)(E(Z_{A|b}) - E(Z_{A|B}))\,, \tag{4.33}$$

$$\frac{dE(Z_{A|b})}{dt} = Ry_B((E(Z_{A|B}) - E(Z_{A|b}))\,, \tag{4.34}$$

$$\frac{dy_B}{dt} = s_2 y_B(1-y_B)\,. \tag{4.35}$$

Similarly, to get second order moments we insert $f = z^2_{A|B}$, $z^2_{A|b}$ and $z_{A|B}z_{A|b}$ and obtain

$$\frac{dE(Z^2_{A|B})}{dt} = E\Big(\frac{Z_{A|B}(1-Z_{A|B})}{2Ny_B} + 2R(1-y_B))Z_{A|B}(Z_{A|b} - Z_{A|B})\Big)\,, \tag{4.36}$$

$$\frac{dE(Z_{A|B}Z_{A|b})}{dt} = R\cdot E\Big((1-y_B)Z_{A|b}(Z_{A|b} - Z_{A|B}) + y_B Z_{A|B}(Z_{A|B} - Z_{A|b})\Big)\,, \tag{4.37}$$

$$\frac{dE(Z^2_{A|b})}{dt} = E\Big(\frac{Z_{A|b}(1-Z_{A|b})}{2N(1-y_B)} + 2Ry_B Z_{A|b}(Z_{A|B} - Z_{A|b})\Big)\,. \tag{4.38}$$

These equations have been introduced by Ohta & Kimura (1975). However, they do not present the derivation from the original gametic frequencies.

Remarkably, the sets of equations for the moments are self contained in the sense that equations for the first moments contain only first order terms and those for the second moments contain only terms of up to order two. This is still true if mutational terms are included. But if selection at locus $\mathcal{A}$ is taken into account, the situation is severely complicated since then equations of order i contain terms of order up to $i+1$. Here, we do not elaborate on the analysis of this more general model and restrict ourselves to the 'classical' hitchhiking situation.

4.3 Effect on Heterozygosity Due to a Single Substitution

To quantify the effect of a single substitution at $\mathcal{B}$, which arises at time $t_1 = 0$ with frequency $y_B(0) = \epsilon$, on heterozygosity at $\mathcal{A}$ we again compute $\bar{H}(t_2)$ at time t_2, when the substitution process is completed ($y_B(t_2) = 1 - \epsilon$). Expected heterozygosity now reads

$$\begin{aligned} E(H(t)) &= 2E\Big\{\Big(y_B(t)Z_{A|B}(t) + (1 - y_B(t))Z_{A|b}(t)\Big) \cdot \\ &\quad \Big(1 - (y_B(t)Z_{A|B}(t) + (1 - y_B(t))Z_{A|b}(t))\Big)\Big\}, \end{aligned} \tag{4.39}$$

which, on expanding the product, obviously involves first *and* second order moments of $Z_{A|B}$ and $Z_{A|b}$. Then, we average over initial conditions to get

$$\bar{E}(H(t)) = Z_{A|b}(t_1) \cdot H(t)|_{Z_{A|B}(t_1)=1} + (1 - Z_{A|b}(t_1)) \cdot H(t)|_{Z_{A|B}(t_1)=0} \cdot \tag{4.40}$$

We know that at time t_2 the quantity y_A may well be approximated (given ϵ is small enough) by $Z_{A|B}$. This can be seen e.g. from (4.39). Heterozygosity therefore is approximately evaluated via $E(Z_{A|B}(1 - Z_{A|B}))$. We subtract (4.36) from (4.33) and obtain

$$\frac{dE(Z_{A|B} - Z_{A|B}^2)}{dt} = -E(Z_{A|B} - Z_{A|B}^2)\left(\frac{1}{2Ny_B} + 2R(1-y_B)\right) + R(1-y_B)E(Z_{A|B} - 2Z_{A|B}Z_{A|b} + Z_{A|b}) \cdot \tag{4.41}$$

We take the weighted average, rewrite $Z_{A|B} - 2Z_{A|B}Z_{A|b} + Z_{A|b}$ as $2(Z_{A|B} - Z_{A|B}Z_{A|b}) - Z_{A|B} + Z_{A|b}$ (the solution for the difference $E(Z_{A|B} - Z_{A|b})$ is readily found by subtracting Eq. (4.34) from Eq. (4.33) to be $E(Z_{A|B} - Z_{A|b})(t) = (Z_{A|B}(t_1) - Z_{A|b}(t_1))e^{-Rt}$) and get

$$\frac{d\bar{E}(Z_{A|B} - Z_{A|B}^2)}{dt} + \bar{E}(Z_{A|B} - Z_{A|B}^2)\left(\frac{1}{2Ny_B} + 2R(1 - y_B)\right) = 2R(1 - y_B)\bar{E}(Z_{A|B} - Z_{A|B}Z_{A|b}) \cdot \tag{4.42}$$

We replace $\bar{E}(Z_{A|B} - Z_{A|B}Z_{A|b})$ by $Z_{A|b}(t_1)(1 - Z_{A|b}(t_1))$. To justify this we note that the mixed second moment $E((Z_{A|B}Z_{A|b})(t))$ is very well approximated by $Z_{A|b}(t_1) \cdot E((Z_{A|B}(t))$ as has been tested by numerical integration. Eq.(4.42) thereby turns into a single linear inhomogeneous differential equation. An approximate solution can be derived in terms of the incomplete Gamma function (see Appendix). This leads to average reduction in heterozygosity of

$$\frac{\bar{H}(t_2)}{\bar{H}(t_1)} = \frac{2R}{s_2}\alpha^{-2R/s_2}\Gamma\left(\frac{-2R}{s_2}, \frac{1}{\alpha}, \frac{1}{\alpha\epsilon}\right). \tag{4.43}$$

$\Gamma(.,.,.)$ denotes the generalized incomplete Gamma function defined by $\Gamma(a,b,c) := \Gamma(a,b) - \Gamma(a,c)$ (formula 8.350.2 in Gradshteyn & Ryzhik (1980)). α is an abbreviation for $2Ns_2$. In Table 4.1 the reduction in heterozygosity due to a single substitution depending on recombination rate R is listed for a choice of parameter values. It is not surprising to find that the effect of a substitution on heterozygosity is estimated to be smaller when the deterministic approach is taken. Stochastic effects are largest as long as frequencies are close to the boundaries 0 or 1. In

$-log_{10}\frac{R}{s_2}$	Eq.(4.16)	Eq.(4.19)	Eq.(4.43)	Numerical	Coalescent
3.0	0.027250	0.027253	0.021631	0.021626	0.021637
2.8	0.042839	0.042847	0.034062	0.034050	0.034079
2.6	0.067033	0.067052	0.053437	0.053409	0.053477
2.4	0.104120	0.104167	0.083341	0.083279	0.083376
2.2	0.159879	0.159989	0.128786	0.128650	0.128580
2.0	0.241173	0.241422	0.196182	0.195883	0.195660
1.8	0.354090	0.354624	0.292346	0.291707	0.291868
1.6	0.499418	0.500456	0.421466	0.420186	0.419731
1.4	0.665391	0.667132	0.579071	0.576761	0.576768
1.2	0.822766	0.825075	0.744880	0.741370	0.740796
1.0	0.934787	0.936904	0.883579	0.879500	0.878784
0.8	0.986376	0.987465	0.965524	0.962341	0.961843
0.6	0.998804	0.999032	0.994531	0.993085	0.992985

Table 4.1: Reduction of heterozygosity due to a single substitution. Parameters: $N = 5 \cdot 10^7$, $\alpha = 2Ns_2 = 10^5$, $\epsilon = 10^{-6}$. 'Numerical' refers to numerical integration of system (4.33) to (4.38) by means of a Runge-Kutta algorithm. The values in the last column ('Coalescent') have been provided by C. H. Langley; they are due to numerical evaluation of the coalescent model for hitchhiking (Kaplan *et al.*, 1989).

$-log_{10}\frac{R}{s_2}$	Simulation				Eq.(4.44)
	$y_A(t_1) = 0.5$		$y_A(t_1) = 0.1$		
	$\alpha = 2 \cdot 10^3$				
3.0	0.0124	(0.0034)	0.0128	(0.0016)	0.0139
2.0	0.1208	(0.0123)	0.1302	(0.0075)	0.1308
1.0	0.7315	(0.0363)	0.7378	(0.0078)	0.7456
0.0	0.9919	(0.0130)	0.9910	(0.0006)	0.9990
	$\alpha = 2 \cdot 10^2$				
3.0	0.0051	(0.0021)	0.0050	(0.0025)	0.0094
2.0	0.0928	(0.0123)	0.0944	(0.0192)	0.0899
1.0	0.5342	(0.0139)	0.5240	(0.0367)	0.5977
0.0	0.9320	(0.0440)	0.9115	(0.0306)	0.9902

Table 4.2: Reduction of heterozygosity due to a single substitution: Comparison of theoretical and simulation results. Parameter: $N = 10^4$. The values in parentheses represent the standard error for the different initial frequencies $y_A(t_1)$. For each parameter set, mean and standard error have been calculated based on 400 simulations.

fact, one observes that deterministic and stochastic approach differ by little, if ϵ is not too small ($\epsilon \geq 5/\alpha$).

For small ϵ ($\epsilon \leq 1/\alpha$) formula (4.43) can be simplified to

$$\frac{\bar{H}(t_2)}{\bar{H}(t_1)} = \frac{2R}{s_2}\alpha^{-2R/s_2}\Gamma(\frac{-2R}{s_2}, \frac{1}{\alpha}). \tag{4.44}$$

We examined this formula by Monte Carlo simulations using a Wright–Fisher model (cf. section 1.5.1). Those selected mutations that go to fixation reduce heterozygosity at the neutral locus by the amounts presented in Table 4.2. The simulation results are compared with Eq.(4.44). They agree well with expected results, if α is sufficiently large, as we have assumed throughout this derivation ('strong' selection). The simulations also indicate that the reduction in expected heterozygosity is independent of the initial frequency of allele A. This is consistent with our approximation result which shows that the right–hand side of Eq.(4.44) is independent of $z_{A|b}(t_1)$.

4.4 Recurring Substitutions: The Long Time Equilibrium

In order to apply the hitchhiking model to evolutionary data, such as variability or divergence (cf. Chapter 5), it is essential to know not only the effect of a single hitchhiking event but

that of recurring events. To model this we assume that substitutions occur at random times according to a Poisson process. We remarked already that two different time scales are involved. Compared to the periods between selective events, evolution during the selective phases proceeds that quickly that we may assume that the process of fixation of strongly selected mutants happens instantaneously. Therefore, at most one selected mutation can be on its way to fixation at any one time. Furthermore, we assume that substitutions ($\mathcal{B}$ locus) occur along the chromosome at random positions with respect to the (fixed) neutral locus $\mathcal{A}$. The distance between the two sites is measured by the recombination rate. We have to be more precise on this point. We let ρ be the recombination rate per nucleotide site per generation. If the physical distance of a selected site from the neutral region ($\mathcal{A}$) is m (measured in base pairs), then its recombinational distance from the neutral region is defined to be ρm, which equals the expected number of cross–overs between the selected site and the neutral region per generation. The relation to the former recombination rate R between the two loci is given by $R = \rho m$. We know that the influence of a substitution on variability at $\mathcal{A}$ becomes negligible, if the (recombinational) distance of the selected site from the neutral region exceeds some maximal value $\bar{R}$. Therefore, we restrict our attention to substitutions which are not farther apart from $\mathcal{A}$ than a maximal recombinational distance $\bar{R}$ or a maximal physical distance $\bar{R}/\rho$ bp. This restriction has another reason. Since we are dealing only with a two–locus model we have to make sure that at any time there is only at most one substitution on its way to fixation.

In determining a maximal value $\bar{R}$ we follow the method outlined by Kaplan *et al.* (1989). See the Appendix for an explicit description of how we computed $\bar{R}$.

To quantify the effect under recurring substitutions we evaluate the probability for the neutral locus to escape *recurring* hitchhiking events. This quantity is the factor by which the neutral value for heterozygosity $H_{neut} := H(0)$ has to be multiplied to yield equilibrium heterozygosity $\bar{H}(\infty)$ 'after' a series of substitutions. We introduce the following notation

ψ the average number of selected substitutions per nucleotide site per generation,

χ the rate at which selected substitutions occur and which eliminate one or the other of the two alleles at $\mathcal{A}$,

$h_{red}(\rho m)$ reduction in heterozygosity due to a single substitution event at distance between m and $m + dm$ from the neutral region.

The latter one corresponds to the right hand side of Eq.(4.43), where R has to be replaced by ρm. dm is a chromosomal stretch of infinitesimal length: We assume the chromosome to be continuous. Furthermore, h_{red} may also be interpreted as the probability with which the neutral locus escapes a *single* hitchhiking event.

The rate at which selected substitutions occur at a distance between m and $m+dm$ and which

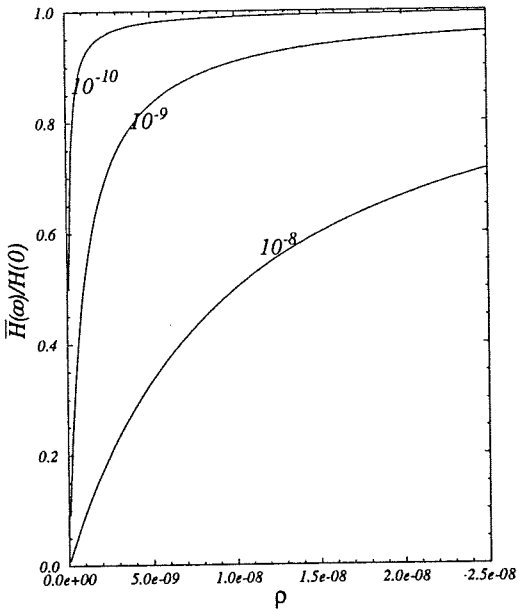

Figure 4.2: Reduction in heterozygosity due to recurring substitutions according to Eq. (4.49). Parameter is $\alpha\psi I_{R^*}(\alpha)$ (cf. Eq. (5.1)). Parameter choices: 10^{-8}, 10^{-9}, 10^{-10}.

eliminate one of the alleles at $\mathcal{A}$ is

$$2N\psi(1 - h_{red}(\rho m))dm\,. \tag{4.45}$$

Thus, the total rate χ is computed to be

$$\chi = 2 \cdot 2N\frac{\psi}{\rho}\int_0^{\bar{R}}(1 - h_{red}(R))dR\,, \tag{4.46}$$

where the integration variable has been scaled back to recombinational distance R. The factor 2 is due to the fact that both sides of the chromosome – left and right with respect to the neutral region – have to be taken into account.

From the rate χ the *probability* that variability at the neutral locus is wiped out under recurring hitchhiking events is computed to be

$$\frac{\chi}{1+\chi}\,. \tag{4.47}$$

Therefore, the probability to escape hitchhiking is

$$\frac{1}{1+\chi}\,. \tag{4.48}$$

Thus, the effect of recurring substitutions on heterozygosity at $\mathcal{A}$ becomes

$$\bar{H}(\infty) = \frac{1}{1+\chi}H_{neut}\,. \tag{4.49}$$

4.5 Summary and Extensions

Above we derived analytical expressions of the effect of a single as well as of recurring substitutions on expected heterozygosity. Reduction of average nucleotide heterozygosity below a level predicted by the neutral theory has been observed by various investigators in natural populations of *Drosophila*, especially in chromosome regions near the centromere and telomere, where recombination is heavily restricted (cf. Chapter 5).

The above considerations are only some first steps of a theory of genetic hitchhiking, because only the effect on neutral polymorphism is taken into account. Given that there is likely to be a spectrum of selected mutations, a more general theory of the evolutionary dynamics of hitchhiking is desirable. It is conceivable that weakly selected alleles interfere with the strongly selected ones and reduce the magnitude of the hitchhiking effect. Similarly, a theory aiming on an understanding of levels of molecular variation needs to include the effects of strongly selected mutations on deleterious alleles, such as alleles carrying transposable elements.

Progress in this direction may be made based on the diffusion approach, since it generalizes readily – other than for example the coalescent approach – if the assumption of neutrality for the hitchhiking alleles is dropped (see Eq. (4.31)).

5

Application: An Estimate on Frequency and Strength of Strongly Selected Substitutions

We have argued that the occurrence and subsequent fixation of strongly selected mutations alters the allelic composition at linked loci. We have also seen this effect to be the stronger the more tight the two loci are linked. We use the relationship between heterozygosity and recombination derived in the previous chapter to estimate some genetical parameters from experimental data.

5.1 Experimental Evidence of Reduced Variation

Studies of genetic variation in *Drosophila* (Aguadé *et al.*, 1989*a*; Stephan & Langley, 1989) show molecular variability to be significantly reduced in certain chromosomal regions where recombination is infrequent (telomeric and centromeric regions). Begun & Aquadro (1992) analyzed available estimates of DNA sequence variation from 20 gene regions in *Drosophila melanogaster*. They demonstrate that nucleotide variation correlates with recombination rate. This observation has also been reported by several other authors (Martín-Campos *et al.*, 1992; Langley *et al.*, 1993; Stephan & Mitchell, 1992; Kindahl & Aquadro, 1995). On the other hand, there is no evidence for an altered nucleotide *mutation* rate in these regions. This has been shown on basis of a statistical test (Hudson *et al.*, 1987) by comparison of inter-specific and intra-specific variation. The level of divergence (i.e. inter-specific variation) is not significantly reduced – unlike the level of diversity (intra-specific variation). If mutation rate were lower in these regions then a simultaneous reduction of both levels had to be expected. As well on grounds of a comparison of inter-specific and intra-specific variation Begun & Aquadro (1992) and Kindahl & Aquadro (1995) reject the

hypothesis of a reduced mutation rate. In this sense, the data contradict the neutral theory. It would predict proportionality of both variation measures, if mutation was the main agent. Rather, some form of selection seems to be responsible for the observed pattern.

In fact, to explain the finding of reduced variation when recombination rate is low the action of hitchhiking may be invoked (Aguadé *et al.*, 1989*a*; Stephan & Langley, 1989; Begun & Aquadro, 1991; Kindahl & Aquadro, 1995). Crucial questions concern the strength and frequency with which selected substitutions have to be postulated in order to explain the observations.

In the last chapter we derived an equation (Eq.(4.49)) which gives equilibrium heterozygosity under recurring substitution events in the neighborhood of a fixed neutral region. We use this equation to obtain an explicit functional relationship between average nucleotide heterozygosity ($\mathcal{H}$) and per nucleotide recombination rate (ρ). To simplify expressions we abbreviate $u := -2R/s_2$ and $R^* := 2N\bar{R}$, where $\bar{R}$ denotes the maximal recombinational distance within which hitchhiking is assumed to be effective. Then, χ can be written as

$$\chi = -\alpha\frac{\psi}{\rho}\int_0^{-2R^*/\alpha}\Big(1 + u\alpha^u\Gamma(u,\frac{1}{\alpha})\Big)du\,, \tag{5.1}$$

with $\alpha = 2Ns_2$. We denote the integral – which does not depend on ρ – by $-I_{R^*}(\alpha)$ and find the function $\mathcal{H} := \bar{H}(\infty)$ to be characterized by (cf. Eq.(4.49))

$$\mathcal{H} = H_{neut}\frac{\rho}{\rho + \alpha\psi I_{R^*}(\alpha)}\,. \tag{5.2}$$

Thus, we may view $\mathcal{H}$ as function of ρ: $\mathcal{H} = \mathcal{H}(\rho)$. However, the equilibrium model of recurring hitchhiking events may lead to misestimates of involved parameters if the recombination rate is extremely small. In this case the assumption of at most one substitution being on its way to fixation might be violated. Furthermore, the recovery of neutral variation due to mutation and drift may not be neglected. We will expand on this point below. For the moment and further analysis we restrict ourselves to the consideration of loci with moderate recombination rates. Our aim is to estimate parameter $\alpha\psi$ in Eq.(5.2), the 'index of selective sweep intensity', using experimental data of nucleotide heterozygosity (commonly referred to as π, defined by Nei & Li (1979)) and per nucleotide recombination rate ρ. For estimation we use two data samples of ρ and π values each. One of them, comprising 20 gene regions on the X– and on autosomal chromosomes in *Drosophila melanogaster*, has been collected by Begun & Aquadro (1992). Sample *II* is due to Kindahl & Aquadro (1995). They kindly provided as yet unpublished variation–recombination data of the third chromosome in *Drosophila melanogaster*. Data are presented in Tables 5.1 and 5.2.

gene region	coefficient of exchange	$\rho \cdot 10^{-9}$	π	Reference
y, ac	0.0045	0.429	0.0008	Begun & Aquadro (1991)
Pgd	0.0154	1.466	0.0030	Begun & Aquadro (1991)
z, tko	0.0222	2.114	0.0044	Aguadé *et al.* (1989*b*)
per	0.0520	4.952	0.0014	Begun & Aquadro (1991)
w	0.1400	13.30	0.0090	Miyashita & Langley (1988)
N	0.1212	11.54	0.0050	Schaeffer *et al.* (1988)
v	0.0590	5.619	0.0010	Begun & Aquadro (1992)
f	0.0455	4.330	0.0020	Langley (1990)
Zw	0.0485	4.619	0.0007	Eanes *et al.* (1989)
$su(f)$	0.0050	0.476	0.0000	Langley (1990)
$Gpdh$	0.0800	5.714	0.0078	Takano *et al.* (1991)
Adh	0.0647	4.621	0.0060	Langley *et al.* (1982)
Ddc	0.0184	1.314	0.0050	Begun & Aquadro (1992)
Amy	0.0435	3.107	0.0080	Langley *et al.* (1988)
Pu	0.0718	5.129	0.0040	Takano *et al.* (1991)
$Est6$	0.0604	4.314	0.0050	Game & Oakeshott (1990)
$MtnA$	0.0083	0.593	0.0010	Lange *et al.* (1990)
$Hsp70A$	0.0069	0.493	0.0020	Leigh-Brown (1983)
ry	0.0471	3.364	0.0030	Aquadro *et al.* (1988)
ci^D	0.0000	0.000	0.0000	Berry *et al.* (1991)

Table 5.1: Coefficient of exchange, recombination rate and nucleotide diversity for different gene regions in *D. melanogaster*. To obtain the recombination rate from the coefficient of exchange the latter has to be multiplied by a factor $2 \cdot 10^{-8}/0.14$ (see text). The first ten of the above loci are X-linked, the last ten are autosomal. For comparability one has to rescale ρ and π values appropriately (see text).

gene region	coefficient of exchange	$\rho \cdot 10^{-9}$	π
$Lsp1\gamma$	0.000	0.000	0.0001
$Hsp26$	0.062	22.44	0.0102
Sod	0.026	9.41	0.0033
$Est6$	0.044	15.93	0.0070
fz	0.016	5.79	0.0043
tra	0.015	5.43	0.0024
Pc	0.002	0.72	0.0010
$Antp$	0.004	1.45	0.0040
Gld	0.004	1.45	0.0022
ry	0.021	7.60	0.0048
Ubx	0.011	3.98	0.0084
$Rh3$	0.033	11.95	0.0061
$E(spl)$	0.051	18.46	0.0070
Tl	0.027	9.77	0.0020
$Mlc2$	0.020	7.24	0.0044

Table 5.2: Coefficient of exchange, recombination rate and nucleotide diversity in third chromosome of *D. melanogaster*. Data from Kindahl & Aquadro (1995).

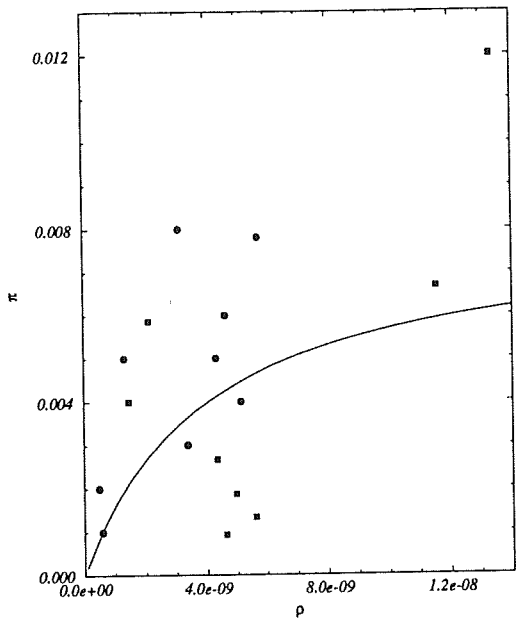

Figure 5.1: Nucleotide diversity vs. recombination rate. Sample *I*. The data for 17 group *I* gene regions are from Begun & Aquadro (1992); loci ci^D, $su(f)$ and y–ac are excluded since their recombination rates are extremely small. Squares: *X*-linked loci; Dots: Autosomal loci. ρ and π values have been scaled (see text).

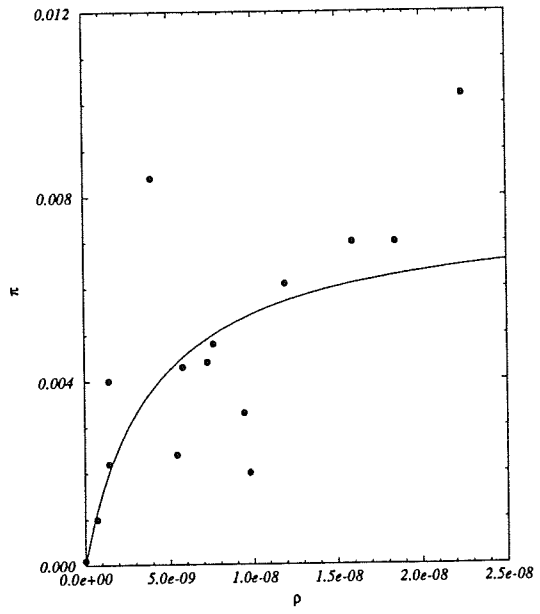

Figure 5.2: Nucleotide diversity vs. recombination rate. Sample *II*. Data are from 15 group *I* genes (see Table 5.2). *Lsp*1γ has been qualified to belong to group *II* and therefore excluded from regression analysis.

5.2 Estimation of Parameters: Fitting the Model

For parameter estimation we identify nucleotide diversity π with (steady state) heterozygosity $\mathcal{H}$ as given in Eq. (5.2). The 'coefficient of exchange' provided by Begun & Aquadro (1992) is a measure which relates the distance (in map units) of flanking regions of the locus under consideration to the number of polytene bands of this region. This measure is – at least to a first approximation – proportional to the recombination rate per nucleotide (ρ) in the gene region of interest. We calculate the proportionality constant for sample I by assuming that the coefficient of exchange for the *white (w)* region corresponds to the average rate of intra-genic recombination in this region. The latter is approximately $2 \cdot 10^{-8}$ per *bp* (Judd, 1987). With this number the proportionality factor becomes approximately $1.43 \cdot 10^{-7}$. Finally, for comparability of the autosomal and X–linked chromosomes one has to rescale the respective data by suitable constants: We multiply ρ values of X–linked loci by the factor 2/3 and those of the autosomal loci by 1/2. Coefficients of exchange are estimated from recombination in females. Since an autosomal gene spends half of the time in males (where there is no recombination in *D. melanogaster*), we multiply by 1/2. X-linked genes spend two-thirds of the time in females where they can recombine and one-third of the time in males where they cannot recombine. Similarly, the π–values of the X–linked loci have to be multiplied by 4/3 to compensate for the difference in effective population size of X-linked versus autosomal genes.

We partition – the reason is explained below – the 20 loci into two groups. Group I comprises the loci with low to high recombination rates, group II those with very low recombination rates, which are ci^D, $su(f)$ and y–ac (the same holds for $Lsp1\gamma$ in sample II).

For sample II it is more appropriate to calibrate the proportionality factor on nucleotide recombination rate of a well studied third chromosome locus. Kindahl & Aquadro (1995) chose the *ry* region and calculated the factor $3.62 \cdot 10^{-7}$.

The model Eq.(5.2) is essentially nonlinear, which triggers the usual problems with parameter estimation (an initial guess for parameter values has to be supplied when applying some least square method; this can prevent any routine from finding reasonable estimates if initial values are bad). A trick from reaction kinetics serves to mend the problem. Performing the so–called Lineweaver–Burk transformation, we just take the reciprocal on both sides of Eq.(5.2) and obtain the linear model

$$\frac{1}{\pi} = \frac{1}{H_{neut}} + \frac{1}{\rho}\frac{\alpha\psi I_{R^\bullet}(\alpha)}{H_{neut}} = c_1 + \frac{1}{\rho}c_2 \,. \tag{5.3}$$

We now estimate parameters c_1 and c_2 by fitting model (5.3) to the reciprocally transformed data. Then one obtains estimates of H_{neut} and $\alpha\psi I_{R^\bullet}(\alpha)$ from those of c_1 and c_2. H_{neut}, the 'neutral' value of heterozygosity, is identified with $4N\mu$, where μ is the average per nucleotide mutation rate (e.g. Crow & Kimura (1970), Chapter 7.2). It is the 'high recombination limit' of average nucleotide heterozygosity and assumed that it characterizes the whole genome rather than

individual loci and that any deviations from H_{neut} are exclusively due to hitchhiking.

For regression analysis we use the so–called *geometric mean* (GM) method. In the terminology of Sokal & Rohlf (1981) this is a 'model *II*' regression procedure, one in which both variables, abscissa and ordinate, are random. The measurements of π *and* the coefficient of exchange are subject to intrinsic measurement errors as well as to stochastic fluctuations due to the evolutionary process. Therefore it seems to be adequate – as the GM method proposes – to regress both variables on each other and to determine a mean value of both regression coefficients. For the GM method, this mean is the geometric mean. Applying GM to the 17 group *I* data of sample *I*, we obtain the following results

$$\begin{aligned} c_1 &= 125.54\,, \\ c_2 &= 5.04 \cdot 10^{-7}\,. \end{aligned}$$

Confidence limits on a 5% level for c_2 are $L_1 = 2.51 \cdot 10^{-7}$ and $L_2 = 7.57 \cdot 10^{-7}$. From these estimates we calculate

$$\begin{aligned} H_{neut} &= 0.0080\,, \\ \alpha\psi I_{R^\bullet}(\alpha) &= 4.03 \cdot 10^{-9}\,. \end{aligned}$$

The analogous estimates for sample *II* are

$$\begin{aligned} c_1 &= 130.67\,, \\ c_2 &= 5.27 \cdot 10^{-7}\,. \end{aligned}$$

Confidence limits (5% level) for c_2 are $L_1 = 3.62 \cdot 10^{-7}$ and $L_2 = 6.92 \cdot 10^{-7}$. The latter estimates lead to

$$\begin{aligned} H_{neut} &= 0.0077\,, \\ \alpha\psi I_{R^\bullet}(\alpha) &= 4.04 \cdot 10^{-9}\,. \end{aligned}$$

As a plot of $I_{R^\bullet}(\alpha)$ vs. α suggests (cf. Figure 5.3), the value of the integral $I_{R^\bullet}(\alpha)$ is nearly independent of α. By interpolation over a range of α from 10^3 to 10^6 we derive

$$I_{R^\bullet} \approx 0.075\,, \tag{5.4}$$

which yields for sample *I*

$$\alpha\psi \approx 5.37 \cdot 10^{-8}\,. \tag{5.5}$$

and for sample *II*

$$\alpha\psi \approx 5.38 \cdot 10^{-8}\,. \tag{5.6}$$

The estimates for both samples agree remarkably well. We analyzed two independent samples to compensate for shortcomings of the applied estimation method, e.g. the GM method does not provide a means to derive confidence intervals for the quantities H_{neut} and $\alpha\psi$.

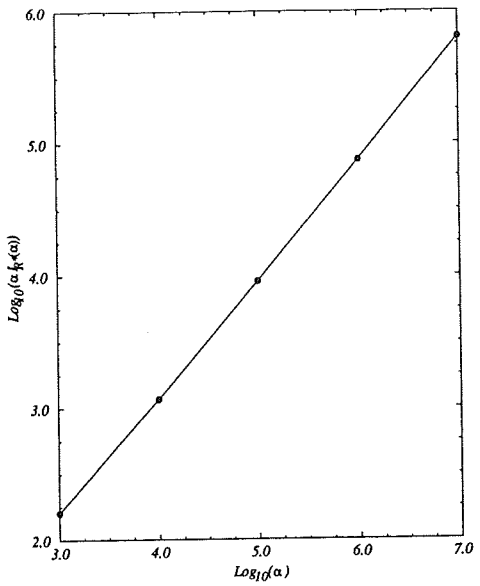

Figure 5.3: $\alpha I_{R^*}(\alpha)$ as function of α. The increase in α is approximately linear.

Using a different argument, it is possible to obtain a lower bound of $\alpha\psi$: H_{neut} is likely to be larger than the level of average nucleotide heterozygosity $\bar{\mathcal{H}}$, obtained by averaging over all gene regions (group I and II), because these include also regions of reduced levels of heterozygosity. Therefore, we have $L_1\bar{\mathcal{H}} < L_1 H_{neut}$. $L_1\bar{\mathcal{H}}$ then is likely to be a lower bound of $\alpha\psi I_{R^*}$. Using the data set from Table 5.1 and correcting of the X–linked loci appropriately, we find $\bar{\mathcal{H}} = 0.004$. Therefore, $\alpha\psi$ is likely to be larger than $1.34 \cdot 10^{-8}$.

So far, we delayed the analysis of hitchhiking in regions of extremely low recombination rates. We will now discuss this point briefly. In regions of *zero* recombination the hitchhiking model would predict *zero* heterozygosity. If this prediction is correct, then any observed variability must be due to mutation and random drift acting between selective events. Recovery of variability due to these forces is also important for very low, non-*zero*, recombination rates. Following Tajima (1989*a*) recovery of heterozygosity between selective sweeps can be described by

$$H(t) = H_{neut} + (H(t_2) - H_{neut}) \exp(-\frac{t - t_2}{2N}), \tag{5.7}$$

where t_2 is the time instant right after a selective sweep. For *zero* recombination we have $H(t_2) = 0$, regardless of the level of heterozygosity before the sweep. Then, to first order, recovery of heterozygosity is given by

$$H(t) = H_{neut}\frac{t - t_2}{2N}, \tag{5.8}$$

if $t - t_2 << 2N$. Rewriting Eq.(5.7) as

$$\frac{H(t) - H(t_2)}{H(t_2)} \approx \left(\frac{H_{neut}}{H(t_2)} - 1\right)\frac{t - t_2}{2N}, \tag{5.9}$$

we note the following. The left hand side denotes the relative departure of heterozygosity at time t from its initial value at time t_2. The right hand side shows that this quantity increases with increasing $H_{neut}/H(t_2)$; but this ratio is larger in regions of low recombination rates. On the other hand, for higher recombination rates this additional contribution to heterozygosity becomes small, since $H_{neut}/H(t_2)$ tends to 1.

5.3 Summary and Extensions

In foregoing sections we applied the model developed in Chapter 4 to experimental data on recombination and variation. We were able to estimate some basic evolutionary parameters by fitting the theoretical model to these data. Although analysis of sample II led to a narrowing of the confidence band with respect to sample I, for more precise estimates additional data are needed. Data in high recombination regions would help to obtain more reliable estimates of H_{neut}. This estimate could be used to obtain also an upper bound for $\alpha\psi$. Interesting appears the question, whether π saturates as recombination increases. If so, one had a direct indicator of a maximal recombination rate beyond which hitchhiking would have no influence on variability of neighboring sites. This in turn could provide a direct measure for average strength of selection per nucleotide. We have seen in the previous chapters that two loci evolve essentially independently iff the recombination rate is of the order of magnitude of the selection coefficient or larger (cf. Figure 4.1).

Furthermore, additional data in low to intermediate domains of recombination rate would reduce the error in the observed π–values. This would be important to delimit the effect of build-up of neutral variation during the recovery phases between hitchhiking events.

So far, as an important result may be stated that the analysis of the steady-state hitchhiking model showed that nucleotide variation, given recombination rate, is characterized essentially by a single parameter, $\alpha\psi$. This result implies that substitutions, as long as they satisfy $\alpha\psi = 2Ns_2\psi = \text{const}$, have the same long term effect on genetic variation. Our analysis appears to imply that the frequency distribution of selection coefficients for strongly selected substitutions follows (perhaps only partially, up to a certain value of s_2) a power law of the form $\psi \propto s_2^{-1}$. Given the probability of fixation for advantageous mutants to be approximately $\approx 2s_2$, this means that the frequency distribution of beneficial mutations is approximately of the form s_2^{-2}. This is a fitness distribution intrinsically different from two others, commonly adopted in the literature. Both are for deleterious mutations and are of the exponential type: One is due to Ohta (1977) (purely exponential), the other to Kimura (1983) (Gamma distribution). These distributions have been introduced mainly for theoretical reasons; it is an open question to what extent they are confirmed experimentally.

With the above results an estimate on population size is quickly derived. Assuming $H_{neut} =$

$4N\mu$ and a reasonable nucleotide mutation rate of $\mu = 10^{-9}$, one calculates

$$N = 2 \cdot 10^6 \,. \tag{5.10}$$

Based on additional assumptions, in particular on the frequency of selected substitutions, the estimates of H_{neut} and $\alpha\psi$ can be used to obtain values for ψ and α (or s_2) separately. There are several ways to proceed, all of which are based on information on sequence variation in regions of *zero* recombination. Sequence analysis of the fourth chromosome locus ci^D revealed that *ten* copies of this gene sampled from a natural *D. melanogaster* population failed to show any variation among 331 silent sites (Berry *et al.*, 1991). Simulating these data, these authors found that there is a 50% probability that an instantaneous selective sweep on the fourth chromosome occured within the last $0.28N$ generations. In the following, we use this result to calculate ψ. Given that the fourth chromosome is approximately *one* percent of the coding genome of *D. melanogaster* and that the haploid genome size is $1.7 \cdot 10^8$bp (Ashburner, 1989), we obtain

$$\psi = \frac{1}{0.28 \cdot 0.01 \cdot 3.4 \cdot 10^8 N} = 1.05 \cdot 10^{-6} N^{-1} \,. \tag{5.11}$$

Inserting our above estimate of N we derive

$$\psi = 5.25 \cdot 10^{-13} \,. \tag{5.12}$$

Using this result, we find α to be approximately $1.02 \cdot 10^5$ and the selection coefficient

$$s_2 = 2.55 \cdot 10^{-2} \,. \tag{5.13}$$

If selected substitutions are relatively rare compared to neutral substitutions, then the proportion of selected substitutions is given by ψ/μ, which evaluates with our estimates to $5.25 \cdot 10^{-4}$. This suggests that the above data can be explained by postulating that *one* in about 2000 substitutions is strongly selected. It is remarkable to find that – in order to explain the data – selective substitutions with a selection coefficient of approximately $2 - 3\%$ have to be postulated. This value is unexpectedly high, given that mutations at the molecular level are thought to have selection coefficients between 10^{-5} and 10^{-3} (Gillespie, 1991).

However, our estimates may be influenced considerably if the reduction in nucleotide heterozygosity is partly due to other forces, such as deleterious mutations. Charlesworth *et al.* (1993) have demonstrated that selection against deleterious alleles may have a similar effect on linked neutral polymorphism as directional selection. If hitchhiking with deleterious alleles plays a role, fewer selected sweeps caused by advantageous mutations and smaller selection coefficients are required to explain the observations.

6

Tying Things Together

Even though over the last four or so years, as molecular data on recombination–variation have been compiled, the hitchhiking model became more widely accepted as a viable explanation of the molecular mechanism leading to the observed data, it has repeatedly been contested on various grounds (Eanes *et al.*, 1989). Braverman *et al.* (1995) only recently reported an observation contradictory to the hitchhiking model, based on the frequency spectrum of segregating nucleotide sites (i.e. point mutations) in population samples. To statistically describe molecular variation in population samples, two statistics are common. One is $\hat{\pi}$ (Nei & Li, 1979), which is the number of nucleotide differences between two sequences averaged over all pairwise comparisons. The other is $\hat{\theta}$, which is the number of segregating sites in a sample, corrected for sample and gene sizes, and is used as estimator for $4N\mu$. Tajima (1989*b*) devised a 'test of neutrality' to test the null hypothesis that observed variation is exclusively due to neutral forces, mutation and drift. This test uses as a principal statistic the difference $D = \hat{\pi} - \hat{\theta}$. Values differing significantly from *zero* should indicate deviation from neutrality. For example, if hitchhiking (a non-neutral agent) was acting, a significantly negative D is expected, since after a selective sweep neutral variation builds up only gradually by mutation and drift. Thus, as Aguadé *et al.* (1989*a*) and Hudson (1990) argue, one would expect that most variants of the wild-type are present at low frequency. $\hat{\pi}$ is sensitive to rare variants, while $\hat{\theta}$ is not: While mutants accumulate after a total elimination of all variation the statistic $\hat{\pi}$ is smaller than $\hat{\theta}$, but $\hat{\pi}$ increases at a higher rate than does the statistic $\hat{\theta}$. Hence, a negative D indicates an excess of rare variants over what would be expected under neutrality, i.e. a situation as after a selective sweep. Now, Braverman *et al.* (1995) stress that for several studies of recombination-variation a significant deviation from *zero* has not been found, which puts doubt on the validity of the hitchhiking model.

It appears to us that in order to give a clearer answer to this problem, the original hitchhiking model has to be refined. We have seen already at the end of the previous chapter, that the dynamics at $\mathcal{A}$ between two hitchhiking events has to be modeled more thoroughly. In particular,

the verbal arguments about excess of rare variants (Aguadé *et al.*, 1989*a*; Hudson, 1990) have to be investigated formally. One possible strategy is obvious. The sequence space model is perfectly adequate for a description of the individual mutation classes. Rather than simplifying the real situation by collecting all alleles except the wild-type into the error tail (as done in the two-locus-two-allele model before), one can keep track of the single mutant classes separately, if necessary. The size $\nu_{\mathcal{A}}$ of the neutral locus under study, is totally neglected in the original hitchhiking model. On the other hand, it is one of the characteristic parameters of the sequence space model. Neutral dynamics between hitchhiking events is readily modeled according to Eq. (1.1). Adopting the view point of negligible back mutations (which is certainly valid for low Hamming classes in case of large $\nu_{\mathcal{A}}$), analytical solutions can successively be obtained for up to an arbitrary number of error classes, since any differential equation for y_i is at most inhomogeneous linear. For example, treating wild-type and *one*-error class separately, we readily compute

$$y_0(t) = \exp(-(1-e^{-\lambda})t)\,, \tag{6.1}$$

$$y_1(t) = \exp(-t(1-e^{-\lambda})-\lambda)\lambda t\,, \tag{6.2}$$

$$y_{et}(t) = 1-\exp(-(1-e^{-\lambda})t)(1+e^{-\lambda}\lambda t)\,, \tag{6.3}$$

where y_{et} denotes the error tail comprising all mutant classes $i \geq 2$ and $\lambda = \nu_{\mathcal{A}}\mu$. In order to see the influence of sequence length we give a simple example. We choose a 'standard' mutation rate of $\mu = 10^{-9}$ and compare sequence lengths $\nu_{\mathcal{A}} = 5\cdot 10^2$ and 10^3 (which is a typical gene size (Berry *et al.*, 1991)). Using the above distribution of wild-type, *one*-error mutants and error tail, one easily computes on a maximum likelihood assumption the expected occupancy of the three classes for a finite sample at any given time. For example, a population sample of size $\hat{N} = 50$ taken at time $t = 10^6$(generations) since the last selective sweep, has most likely the composition as given in the following table. We see that 1) the wild-type is expected to be definitely more

sample size	$\nu_{\mathcal{A}}$	number of sample in class		
		wild-type	*one*-error mutants	error tail
50	500	31	15	4
50	1000	19	18	13

frequent in the first case and that 2) occupancy of the *one*-error class is expected to be about the same in both cases. Both these observations support that $\hat{\pi}$ and $\hat{\theta}$ tend to be smaller if smaller chromosome regions are surveyed for variability. If the population contains many wild-type alleles then one rarely detects segregating nucleotides.

Just comparing $y_0(t)$ and $1-y_0(t)$ for different chain lengths and reasonably assuming that ψ, the rate of selected substitutions, is the same, then it becomes clear that, at any time, it will be more likely to collect a more heterogeneous sample if $\nu_{\mathcal{A}}$ is larger, therefore $\hat{\pi}$ and $\hat{\theta}$ are

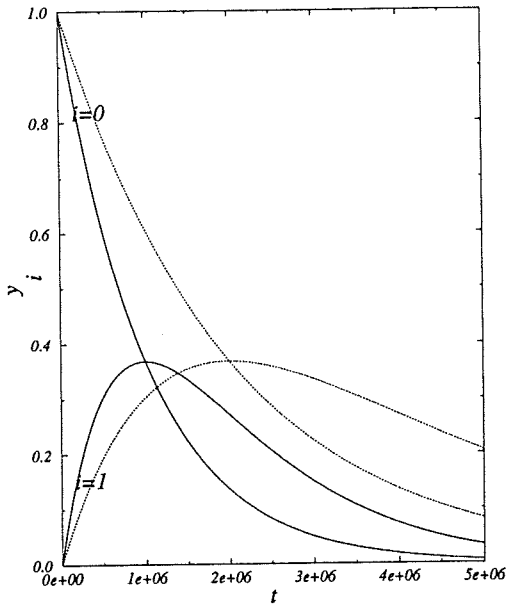

Figure 6.1: Recovery of diversity under neutrality (i.e. all fitness values identical 1) due to mutation after a selective sweep eliminated all variation. Plotted are frequencies of the new wild-type (which had been fixed by hitchhiking), $i = 0$, and *one*-error class, $i = 1$. Parameters: $\mu = 10^{-9}$; $\nu_{\mathcal{A}} = 5 \cdot 10^2$ (dotted) and $\nu_{\mathcal{A}} = 10^3$ (solid).

expected to be larger as well (cf. Figure 6.1). But is the increase of these two measures different for different gene sizes? To preliminary answer this question we carried out computer simulations and calculated both values for a population sample. An initially homogeneous population of size N evolved under the influence of drift and mutation according to a time–homogeneous birth–death process, where selection has been absent (all individuals have the same fitness). Sequences are modeled as binary strings as described in Chapter 1. We collected a sample of size $\hat{N}$ at generation t_1 and calculated $\hat{\pi}$ and $\hat{\theta}$ according to

$$\hat{\pi} = \frac{1}{\hat{N}^2} \sum_{i,j=1}^{\hat{N}} \frac{\pi_{ij}}{\nu_{\mathcal{A}}} \tag{6.4}$$

and

$$\hat{\theta} = \frac{S}{\nu_{\mathcal{A}} \log(\hat{N})}, \tag{6.5}$$

where S denotes the number of segregating sites in the sample and π_{ij} the number of differing sites in a pairwise comparison of string i and j. The results suggest that the difference $\hat{\pi} - \hat{\theta}$ tends to be larger for larger gene sizes $\nu_{\mathcal{A}}$. However, for smaller mutation rates variances increase dramatically. On the other hand, D values did in our simulations not show a trend correlated with population size N. These observations imply that one has to be aware of the possibility that non-significant deviations of D from *zero* may still be compatible with the assumption that hitchhiking is responsible for the observed pattern of variation.

Parameters			$\hat{\pi}$	$\hat{\theta}$	$D = \hat{\pi} - \hat{\theta}$	
ν_A	t_1	μ				
50	10^2	10^{-4}	$1.911 \cdot 10^{-2}$	$9.824 \cdot 10^{-2}$	$-7.913 \cdot 10^{-2}$	$(2.680 \cdot 10^{-3})$
100	10^2	10^{-4}	$1.923 \cdot 10^{-2}$	$9.863 \cdot 10^{-2}$	$-7.940 \cdot 10^{-2}$	$(1.917 \cdot 10^{-3})$
50	10^2	10^{-6}	$2.249 \cdot 10^{-4}$	$1.431 \cdot 10^{-3}$	$-1.207 \cdot 10^{-3}$	$(5.723 \cdot 10^{-4})$
100	10^2	10^{-6}	$2.391 \cdot 10^{-4}$	$1.521 \cdot 10^{-3}$	$-1.282 \cdot 10^{-3}$	$(2.529 \cdot 10^{-4})$
500	10^6	10^{-9}	$5.053 \cdot 10^{-6}$	$2.702 \cdot 10^{-5}$	$-2.197 \cdot 10^{-5}$	$(4.811 \cdot 10^{-5})$
1000	10^6	10^{-9}	$4.628 \cdot 10^{-5}$	$9.294 \cdot 10^{-5}$	$-4.665 \cdot 10^{-5}$	$(7.121 \cdot 10^{-5})$

Table 6.1: Simulation results for quantities $\hat{\pi}$ and $\hat{\theta}$. We simulated neutral evolution of an initially homogeneous population according to an algorithm described by Gillespie (1976). $\hat{\pi}$ and $\hat{\theta}$ values are obtained as averages from 14 simulation runs (unbiased estimators of standard deviation in parentheses), in each of which we averaged over 100 samples each of size $\hat{N} = 50$. Total population size $N = 10^5$. Help with implementation and adaptation of the simulation routines by Walter Grüner is gratefully acknowledged.

Another argument for a sequence space approach to refine the hitchhiking model is the following. Braverman *et al.* (1995) criticize that current ones do not provide distributional properties of their statistics and only give expectations or variances (cf. Eq.s (4.33) to (4.38)). As soon as it comes to getting a more detailed picture of the actual frequency distribution of segregating sites (i.e., equivalently, of error class composition), a treatment of the dynamics in terms of the sequence space model becomes indispensable. Doing this, one might be able also to answer the question about the evolutionary role of hitchhiking more clearly.

Our investigations originated from the general sequence space model, where we observed that the interplay of mutation and selection led to some interesting features of the stationary frequency distribution of a population of nucleotide strings. The original model has been conceived for description of prebiotic evolution, but we have seen it to carry over directly into the biotic context of evolution of natural organisms. With a reinterpretation of parameters replication-mutation systems are equivalent to Fisher's well known mutation–selection equation for multiple alleles. We introduced the concept of recombination which enabled us to describe evolution of a gene site more realistically, since it obviously depends not only on its own selective regime, but also on that of neighboring sites. Only by recombination decoupling of adjacent sites is possible which opens the possibility of independent evolution of different sites and thus becomes a means to avoid the error catastrophe. Specialization of the two locus two allele model led to the classical hitchhiking scenario. We discussed – using the diffusion approach – how hitchhiking affects equilibrium heterozygosity in finite populations. Based on this model we were able to derive estimates for some evolutionary parameters by comparison with experimental data relating nucleotide diversity and recombination.

Appendix

Claim A.1 (p. 32) *Let M* (see (1.11)) *denote the $(\nu+1)\times(\nu+1)$ mutation matrix of the model with backflow and $\hat{M}$* (see (2.3)) *the restriction of the originally infinitely large mutation matrix of the model without backflow to size $(\nu+1)\times(\nu+1)$. Then, for any fixed i and j, m_{ij} and $\hat{m}_{ij}$ are approximately identical as ν becomes large (and as p becomes small at the same time).*

PROOF. We have

$$m_{ij} = (1-p)^{\nu} \sum_{k=j+i-\nu}^{\min(i,j)} \binom{j}{k}\binom{\nu-j}{i-k}\epsilon^{i+j-2k}$$

with $\epsilon := \frac{p}{1-p}$ and

$$\hat{m}_{ij} = \begin{cases} \exp(-\lambda)\frac{\lambda^{i-j}}{(i-j)!} & , \quad i \geq j \\ 0 & , \quad i < j \end{cases},$$

where we identify $\lambda = \nu p$. The proof goes by induction on the mutation order n, i.e. the number of mutating nucleotides, for each entry. For instance, order $n = 0$ gives the probability to copy all nucleotides accurately. Indices i, j take values from 0 to ν.

$n = 0$:

The only non trivial entries are on the main diagonal. $\hat{m}_{ij} = \exp(-\lambda)$, if $i = j$ and *zero* elsewhere. The mutation order being *zero*, the exponent of ϵ, $i + j - 2k$, has to be equal *zero*. This leads to $k = (i + j)/2$. Thus, $m^{(0)}_{ij} = (1-p)^{\nu}$ in case $i = j$ and $m^{(0)}_{ij} = 0$ elsewhere. But $(1-p)^{\nu} \approx \exp(-\lambda)$ for p small and ν large.

$n = 1$:

Non trivial entries are to be found on the main- and first sub- diagonal in $\hat{M}$ and additionally on the first superdiagonal in M. Entries on the main diagonal are the same as in case order 0. Entries on the first subdiagonal are: $m^{(1)}_{i+1,i} = (1-p)^{\nu}(\nu - i)\epsilon^1 = p\nu(1-p)^{\nu-1} - pi(1-p)^{\nu-1} = (p(\nu-1)-p(i-1))(1-p)^{\nu-1} \approx (\lambda - O(1/\nu))\exp(-\lambda) \approx \hat{m}_{i+1,i}$ (i is kept fixed!). Entries on the first superdiagonal are: $m^{(1)}_{i,i+1} = (1-p)^{\nu}(i+1)\epsilon^1 = p(i+1)(1-p)^{\nu-1} = O(1/\nu)\exp(-\lambda) \approx \hat{m}_{i,i+1} = 0$.

$(n-1) \to n$ (n even):

Put $d := (i-j)/2$. Then d ranges from $-\frac{n}{2}$ to $\frac{n}{2}$. The reason for this observation is the following. In order to see which new non trivial entries have been introduced by increasing the mutation order, one has to look only at those i, j such that $\frac{i+j-n}{2} \leq j$ (since have to fulfill $\binom{j}{k} > 0$), which leads to $\frac{i-j}{2} \overset{!}{\leq} \frac{n}{2}$. Furthermore, $i-k$ has to be positive (since want $\binom{\nu-j}{i-k} > 0$), which leads to $i - \frac{i+j-n}{2} \geq 0$ or $\frac{i-j}{2} \geq -\frac{n}{2}$. Now, for the n-th mutation order, we have $m_{ij}^{(n)} = (1-p)^{\nu}\binom{j}{k}\binom{\nu-j}{i-k}\epsilon^n$, where $k = \frac{i+j-n}{2}$ ($i+j$ has to be even for $m_{ij}^{(n)}$ not to be trivial). In terms of d, $m_{ij}^{(n)}$ becomes $(1-p)^{(\nu-n)}p^n\binom{j}{j+d-\frac{n}{2}}\binom{\nu-j}{\frac{n}{2}+d}$. Now, we consider two cases.

1) $d = \frac{n}{2}$, i.e. $i-j = n$: The above expression becomes

$$
\begin{aligned}
&(1-p)^{\nu-n}p^n\binom{\nu-j}{n} = (1-p)^{\nu-n}(p\nu)^n\frac{(\nu-j)!}{n!(\nu-j-n)!\nu^n} \\
&= (1-p)^{\nu-n}\frac{\lambda^n}{n!}\frac{(\nu-j)\ldots(\nu-j-n+1)}{\nu^n} \\
&= (1-p)^{\nu-n}\frac{\lambda^n}{n!}(1-\frac{j}{\nu})\ldots(1-\frac{j+n-1}{\nu}) \; .
\end{aligned}
$$

Expansion of the terms in the last n parentheses leads to an expression equal to $1+O(1/\nu)$. Thus, the whole expression is asymptotically identical to $\exp(-\lambda)\frac{\lambda^n}{n!}$, as required.

2) $-\frac{n}{2} \leq d < \frac{n}{2}$: Now, we have to look at the asymptotic behavior of

$$
\begin{aligned}
&(1-p)^{\nu-n}p^n\binom{j}{j+d-\frac{n}{2}}\binom{\nu-j}{\frac{n}{2}+d} \\
&= (1-p)^{\nu-n}\lambda^n\frac{j!(\nu-j)!}{(j+d-\frac{n}{2})!(\frac{n}{2}-d)!(\frac{n}{2}+d)!(\nu-j-\frac{n}{2}-d)!\nu^n} \\
&= (1-p)^{\nu-n}\lambda^n\frac{j!}{(j+d-\frac{n}{2})!(\frac{n}{2}-d)!(\frac{n}{2}+d)!}\frac{(\nu-j)\ldots(\nu-j-\frac{n}{2}-d+1)}{\nu^n} \; .
\end{aligned}
$$

The enumerator of the last fraction contains only $n/2+d$ terms, which is strictly less than n. Thus, in fact we have an expression of order $O(1/\nu)$, forcing the whole term to have the asymptotical value 0.

Thus, through addition of mutation order n, essential terms are introduced only on the n-th subdiagonal. Any other entries remain 0 (terms below the n-th subdiagonal) or do at least not change their asymptotic behavior, so as already determined by lower mutation orders (terms above the n-th subdiagonal).

$(n-1) \to n$ (n odd):

Analogous to case n even. □

Claim A.2 (p. 33) *Let* $y_i = \binom{\nu}{i}(\frac{p}{s})^i(1-\frac{p}{s})^{\nu-i}$, $v_i = (1-s)^i$ *and* $m_{ij} = \binom{\nu-j}{i-j}p^{i-j}(1-p)^{\nu-i}$. *Then* y_i *satisfy*

$$\sum_{j=0}^{i} y_j v_j m_{ij} = y_i \bar{v} \,. \tag{A.1}$$

Lemma

$$\sum_{i=0}^{\nu} \binom{\nu}{i} (\frac{p}{s})^i (1-\frac{p}{s})^{\nu-i}(1-s)^i = (1-p)^\nu \,. \tag{A.2}$$

PROOF of lemma. By induction on ν. $\nu = 0$: The assertion is true. $\nu \to \nu+1$:

$$\sum_{i=0}^{\nu+1} \binom{\nu+1}{i} (\frac{p}{s})^i (1-\frac{p}{s})^{\nu+1-i}(1-s)^i =$$

$$(1-\frac{p}{s}) \sum_{i=0}^{\nu} \binom{\nu}{i} (\frac{p}{s})^i (1-\frac{p}{s})^{\nu-i}(1-s)^i + \frac{p}{s} \sum_{i=0}^{\nu-1} \binom{\nu}{i} (\frac{p}{s})^i (1-\frac{p}{s})^{\nu-i}(1-s)^{i+1} + (\frac{(1-s)p}{s})^{\nu+1} =$$

$$(1-\frac{p}{s})(1-p)^\nu + \frac{(1-s)p}{s}\Big((1-p)^\nu - (\frac{(1-s)p}{s})^\nu\Big) + (\frac{(1-s)p}{s})^{\nu+1} =$$

$$(1-p)^\nu(1-\frac{p}{s}) + (1-p)^\nu \frac{(1-s)p}{s} = (1-p)^{\nu+1} \,.$$

□

PROOF. By the previous lemma follows $\bar{v} = (1-p)^\nu$. Then, by straightforward calculation, we compute

$$\sum_{j=0}^{i} y_j v_j m_{ij} =$$

$$\ldots = (\frac{s-p}{s})^\nu p^i (1-p)^{\nu-i} \binom{\nu}{i} \sum_{j=0}^{i} \binom{i}{j} (\frac{1-s}{s-p})^j =$$

$$\ldots = \binom{\nu}{i} (1-p)^\nu (\frac{s-p}{s})^\nu (\frac{p}{s-p})^i = \ldots = y_i (1-p)^\nu \,.$$

□

Claim A.3 (p. 33) *For landscape* L_{SP} *and the mutation matrix as in Eq.(2.2) hold*

$$y_0 = \frac{(1-p)^\nu - (1-s)}{s} \,, \tag{A.3}$$

and

$$E(p) = \frac{\nu p (1-p)^{\nu-1}}{(1-p)^{\nu-1} - (1-s)} \,. \tag{A.4}$$

PROOF. The first part follows immediately from the equation for the master class

$$y_0 v_0 (1-p)^\nu = y_0 \bar{v}$$

and the fact that for the single peaked landscape

$$\bar{v} = 1 - s + sy_0 \,.$$

As long as $y_0 \neq 0$ one has $\bar{v} = (1-p)^\nu$. To calculate $E(p)$, we multiply the equilibrium equation on both sides by k, then sum over k to obtain

$$\sum_{k=0}^{\nu}\sum_{i=0}^{k} k y_i (1-s)\binom{\nu-i}{k-i} p^{k-i}(1-p)^{\nu-k} + sy_0\nu p = E(p)\cdot(1-p)^\nu \,.$$

On changing the order of summation and readjusting summation indices the latter equation becomes

$$(1-s)\Big(\sum_{i=0}^{\nu} y_i \sum_{k=0}^{\nu-i} k\binom{\nu-i}{k} p^k (1-p)^{\nu-i-k} + \sum_{i=0}^{\nu} i y_i \sum_{k=0}^{\nu-i}\binom{\nu-i}{k} p^k (1-p)^{\nu-i-k}\Big) + sy_0\nu p =$$

$$(1-s)\Big(\sum_{i=0}^{\nu} y_i(\nu-i)p + \sum_{i=0}^{\nu} i y_i\Big) + \Big((1-p)^\nu - (1-s)\Big)\nu p =$$

$$(1-s)\Big(\nu p + E(p)\cdot(1-p)\Big) + \Big((1-p)^\nu - (1-s)\Big)\nu p =$$

$$E(p)\cdot(1-p)^\nu \,.$$

Solving the last equation for $E(p)$ the required expression emerges. □

Claim A.4 (p. 64) *Let $c \geq 0$. Then*

$$\lim_{n\to\infty} e^{-cn}\cdot\sum_{k=0}^{n}\frac{(cn)^k}{k!} = \begin{cases} 0 & \text{if } c > 1 \\ 1 & \text{if } c < 1. \end{cases} \tag{A.5}$$

PROOF. Put $e^{-cn}\cdot\sum_{k=0}^{n}\frac{(cn)^k}{k!} = f_n(c)$ and $\lim_{n\to\infty} f_n(c) = f(c)$. Then

$$\frac{d}{dc}f(c) = -\delta_1(c)\,, \tag{A.6}$$

where $\delta_1(c)$ is the Dirac delta distribution with its peak at $c = 1$. This can be seen as follows (cf. Figure A.1). We have

$$\frac{d}{dc}f_n(c) = -e^{-cn}\frac{(cn)^n}{(n-1)!}\,. \tag{A.7}$$

Replacing the factorial in the denominator according to Stirling's formula by

$$(\frac{n-1}{e})^{n-1}\cdot\sqrt{2\pi(n-1)} \tag{A.8}$$

we derive (for n sufficiently large)

$$\frac{d}{dc}f_n(c) = \frac{\exp(n\log\frac{cn}{n-1} + n(1-c) - 1 + \log(n-1))}{\sqrt{2\pi(n-1)}}\,. \tag{A.9}$$

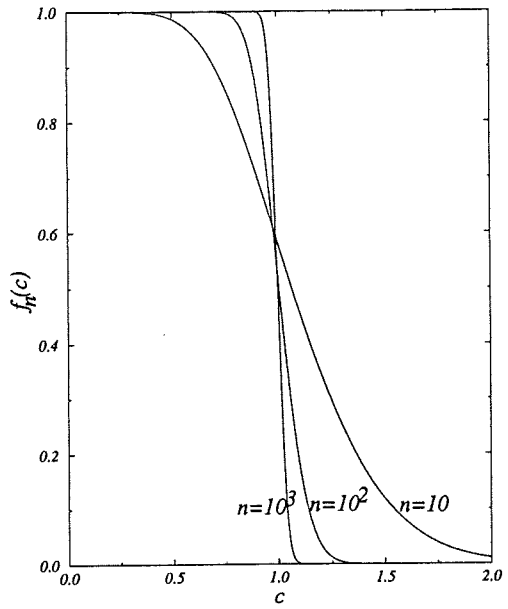

Figure A.1: Plot of $f_n(c)$ for various values of n. The limit for large n is a step function. Parameters as before.

Taking the pointwise limit, one derives

$$\lim_{n\to\infty} \frac{d}{dc} f_n(c) = \begin{cases} 0 & \text{if } c \neq 1 \\ -\infty & \text{if } c = 1. \end{cases} \tag{A.10}$$

Furthermore, we note that

$$\begin{aligned} \lim_{n\to\infty} f_n(0) &= 1 \text{ and} \\ \lim_{c\to\infty} \lim_{n\to\infty} f_n(c) &= \lim_{n\to\infty} \lim_{c\to\infty} f_n(c) = 0\,. \end{aligned}$$

This, together with the information on the behavior of the derivative, proves the claim. □

Claim A.5 (p. 74)

$$\frac{d}{dt} E(f,t) = E(\mathcal{L}(f),t)\,. \tag{A.11}$$

We give the proof for the special case that the (space-)dimension equals *one*. It generalizes to the case of space dimension $= n$ by use of analogous arguments.

PROOF. The right side reads

$$E(\mathcal{L}(f),t) = E\Big(\frac{1}{2} a_{11}(x) \frac{\partial d^2}{\partial x^2} f(x,t) + b_1(x) \frac{\partial d}{\partial x} f(x,t), t\Big)\,.$$

With the density $\phi(x,t)$ the operator E has the meaning

$$E(f,t) = \int_0^1 f(x)\phi(x,t)dx\,.$$

According to the law of total probability, we have

$$\phi(x, t_1 + t_2) = \int_0^1 \text{Prob}(x, t_1 + t_2 | \xi, t) \phi(\xi, t_1) d\xi \,,$$

where $\text{Prob}(x, t_1 + t_2 | \xi, t_1)$ is the conditional probability that the process is in state x at time $t_1 + t_2$, given it has been in state ξ at time t_1 ($t_1 > 0, t_2 \geq 0$). We expand f in a Taylor series, retain terms of up to order *two* and get

$$\begin{aligned}
&E(f, t+h) = \\
&\int_0^1 f(x)\phi(x, t+h)dx = \\
&\int_0^1 f(x) \int_0^1 \text{Prob}(x, t+h|\xi, t)d\xi dx = \\
&\int_0^1 \Big(\int_0^1 f(\xi)\text{Prob}(x, t+h|\xi, t)dx + \int_0^1 f'(\xi)(x-\xi)\text{Prob}(x, t+h|\xi, t)dx + \\
&\frac{1}{2}\int_0^1 f''(\xi)(x-\xi)^2\text{Prob}(x, t+h|\xi, t)dx\Big)\phi(\xi, t)d\xi \,.
\end{aligned}$$

Furthermore,

$$E(f, t) = \int_0^1 \Big(\int_0^1 f(\xi)\text{Prob}(x, t+h|\xi, t)dx\Big)\phi(\xi, t)d\xi \,.$$

Subtracting the last equation from the previous one and dividing by h, we obtain

$$\begin{aligned}
&\frac{1}{h}\Big(E(f, t+h) - E(f, t)\Big) = \\
&E\Big(f' \int_0^1 \frac{1}{h}(x-\xi)\text{Prob}(x, t+h|\xi, t)dx, t\Big) + E\Big(\frac{1}{2}f'' \int_0^1 \frac{1}{h}(x-\xi)^2\text{Prob}(x, t+h|\xi, t)dx, t\Big) \,.
\end{aligned}$$

Assuming that

$$\lim_{h\to 0} \int_0^1 \frac{1}{h}(x-\xi)\text{Prob}(x, t+h|\xi, t)dx = b_1(\xi, t)$$

and

$$\lim_{h\to 0} \int_0^1 \frac{1}{h}(x-\xi)^2\text{Prob}(x, t+h|\xi, t)dx = a_{11}(\xi, t) \,,$$

the above difference leads – when passing to the limit ($h \to 0$) – to the asserted equality. □

Claim A.6 (p. 75) *The inhomogeneous linear differential equation*

$$\frac{d}{dt}f(t) + f(t)h(t) = g(t) \tag{A.12}$$

with

$$\begin{aligned}
f(t) &= \bar{E}(Z_{A|B}(t) - Z^2_{A|B}(t))/(Z_{A|b}(t_1) - Z^2_{A|b}(t_1)) \\
h(t) &= \frac{1}{2Ny_B(t)} + 2R(1 - y_B(t)) \\
g(t) &= 2R(1 - y_B(t))
\end{aligned}$$

evaluates at $t = t_2$ approximately to

$$f(t_2) \approx \frac{2R}{s_2}\alpha^{-2R/s_2}\Big(\Gamma(-\frac{2R}{s_2}, \frac{1}{\alpha}) - \Gamma(\frac{2R}{s_2}, \frac{1}{\alpha\epsilon})\Big)\,. \tag{A.13}$$

PROOF. From the general solution of the homogeneous equation

$$f_{hom}(t) = c\exp(-\int_0^t h(\tau)d\tau)$$

one formally obtains the particular solution of the inhomogeneous equation (satisfying the initial condition $f(0) = 0$) by variation of constants

$$f(t) = c(t)f_{hom}(t)\,,$$

where $c(t) = \int_0^t g(\tau)\exp(\int_0^\tau h(\tau')d\tau')d\tau$. The integral in the former equation evaluates to

$$\int_0^t h(\tau)d\tau \approx \frac{1}{\alpha\epsilon}(1 - e^{-s_2 t}) - \frac{2R}{s_2}\log(\epsilon + e^{-s_2 t})$$

and furthermore, for $t = t_2$,

$$\exp(-\int_0^t h(\tau)d\tau) \approx \epsilon^{\frac{2R}{s_2}} e^{-\frac{1}{\alpha\epsilon}}\,. \tag{A.14}$$

Then, we write

$$c(t_2) = 2R\exp(\frac{1}{\alpha\epsilon})\int_0^{t_2} \frac{e^{-s_2 t}}{\epsilon + e^{-s_2 t}}\exp\Big(-\frac{1}{\alpha\epsilon}e^{-s_2 t} - \frac{2R}{s_2}\log(\epsilon + e^{-s_2 t})\Big)dt$$

and note that $c(t)$ has a form similar to $I(t)$ in Eq.(4.17). We use the same approximation technique and integrate only from 0 to $t_2/2$ instead of t_2. Then $\epsilon + e^{-s_2 t} \approx e^{-s_2 t}$ and $c(t_2)$ simplifies to

$$c(t_2) \approx 2R\exp(\frac{1}{\alpha\epsilon})\int_0^{t_2/2} \exp(-\frac{1}{\alpha\epsilon}e^{-s_2 t} + 2Rt)dt\,.$$

The latter integral can be written by means of the incomplete Gamma function. Together with Eq.(A.14) one obtains the asserted expression Eq.(A.13). □

Remark (p. 78,82) To determine maximal recombination rates $R^* := 2N\bar{R}$ within which hitchhiking is assumed to be effective, we took the following strategy, which had also been adopted in the coalescent hitchhiking model by Kaplan *et al.* (1989). On the one hand, one has to take a neighborhood as large as possible around the neutral region into account for hitchhiking, on the other hand, this region has to be limited in order not to violate the assumption of at most one substituting allele being on its way to fixation. We do this by postulating

$$\int_0^{-2R^*/\alpha}(1 + u\alpha^u\Gamma(u, \frac{1}{\alpha}))du = \frac{1}{1+\delta}\int_0^{-2R_f^*/\alpha}(1 + u\alpha^u\Gamma(u, \frac{1}{\alpha}))du\,, \tag{A.15}$$

where δ is a small positive constant. Now, we choose R_f^* as the point where the integrand function (as function of u) in the above expression is 1% away from its asymptotic limit. In this way we

	R^*		$\alpha I_{R^*}(\alpha)$	
α	$\delta = 0.01$	$\delta = 0.05$	$\delta = 0.01$	$\delta = 0.05$
10^3	301	235	160.5	154.4
10^4	2,396	1,756	1,164.0	1,119.7
10^5	18,696	13,594	9,103.2	8,756.4
10^6	152,350	110,845	74,806.3	71,956.5

Table A.1: Numerical values of R^* and $\alpha I_{R^*}(\alpha)$ according to condition Eq. (A.15).

make sure that the integral on the left hand side of the above equation changes only by a small and definable amount δ as R^* is increased to R_f^*. We computed the following numerical values for R^* (see Table A.1), which depend only on α, but not on N or s_2 separately. Knowing the upper integration limits, numerical values for the integral $I_{R^*}(\alpha)$ in Eq.(5.1) can be determined as well.

Bibliography

AGUADÉ, M., MIYASHITA, N. & LANGLEY, C. (1989*a*). Reduced variation in the *yellow-achaete-scute* region in natural populations of *Drosophila melanogaster*. *Genetics* **122**, 607–615.

AGUADÉ, M., MIYASHITA, N. & LANGLEY, C. (1989*b*). Restriction map variation at the *zeste-tko* region in natural populations of *Drosophila melanogaster*. *Mol. Biol. Evol.* **6**, 123–130.

AKIN, E. (1979). *The Geometry of Population Genetics*, volume 31 of *Lect. Notes Biomath.* New York: Springer.

AQUADRO, C. F., LADO, K. M. & NOON, W. A. (1988). The *rosy* region of *Drosophila melanogaster* and *Drosophila simulans*. I. Contrasting levels of naturally occuring DNA restriction map variation and divergence. *Genetics* **119**, 875–888.

ASHBURNER, M. (1989). *Drosophila: A Laboratory Handbook*. Cold Spring Harbor: Cold Spring Harbor Press.

BAAKE, E. & WIEHE, T. Bifurcations in diploid models on sequence space. J. Math. Biol. (submitted), (1995).

BEGUN, D. & AQUADRO, C. F. (1991). Molecular population genetics of the distal portion of the *X* chromosome in *Drosophila*: Evidence for genetic hitchhiking of the *yellow-achaete* region. *Genetics* **129**, 1147–1158.

BEGUN, D. & AQUADRO, C. F. (1992). Levels of naturally occuring DNA polymorphism are correlated with recombination rates in *Drosophila melanogaster*. *Nature* **356**, 519–520.

BERRY, A. J., AJIOKA, J. W. & KREITMAN, M. (1991). Lack of polymorphism on the *Drosophila* fourth chromosome resulting from selection. *Genetics* **129**, 1111–1117.

BRAVERMAN, J. M., HUDSON, R. R., KAPLAN, N. L., LANGLEY, C. H. & STEPHAN, W. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics (in press), (1995).

BÜRGER, R. (1983). On the evolution of dominance modifiers I. A nonlinear analysis. *J. Theor. Biol.* **101**, 585–598.

CHAO, L. (1988). Evolution of sex in RNA viruses. *J. Theor. Biol.* **133**, 99–112.

CHARLESWORTH, B., MORGAN, M. T. & CHARLESWORTH, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289–1303.

CROW, J. F. & KIMURA, M. (1970). *An Introduction to Population Genetics Theory.* New York: Harper and Row.

DOBZHANSKY, T. (1955). A review of some fundamental concepts and problems of population genetics. *Cold Spring Harbor Symp. Quant. Biol.* **20**, 1–15.

EANES, W. F., AJIOKA, J. W., HEY, J. & WESLEY, C. (1989). Restriction map variation associated with the *G6pd* polymorphism in natural populations of *Drosophila melanogaster. Mol. Biol. Evol.* **6**, 384–397.

EIGEN, M. & SCHUSTER, P. (1977). The hypercycle A: A principle of natural self-organization: Emergence of the hypercycle. *Naturwissenschaften* **64**, 541–565.

EIGEN, M., MCCASKILL, J. & SCHUSTER, P. (1989). The molecular quasi–species. *Adv. Chem. Phys.* **75**, 149–263.

EIGEN, M. (1971). Selforganization of matter and the evolution of biological macromolecules. *Naturwiss.* **58**, 465–523.

EIGEN, M. (1993/7). Viral quasispecies. *Scientific American* pages 32–39.

ERDELYI, A. (1953). *Higher Transcendental Functions*, volume 1. New York: McGraw-Hill.

ETHIER, S. N. & KURTZ, T. G. (1986). *Markov Processes.* New York: John Wiley & Sons.

ETHIER, S. N. & NAGYLAKI, T. (1989). Diffusion approximations of the two-locus Wright-Fisher model. *J. Math. Biol.* **27**, 17–28.

EWENS, W. J. (1979). *Mathematical Population Genetics.* Berlin: Springer.

FONTANA, W. & SCHUSTER, P. (1987). A computer model of evolutionary optimization. *Biophys. Chem.* **26**, 123–147.

GAME, A. Y. & OAKESHOTT, J. G. (1990). The association between restriction site polymorphism and enzyme activity variation for *Esterase6* in *Drosophila melanogaster. Genetics* **126**, 1021–1031.

GILLESPIE, D. T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comp. Phys.* **22**, 403–434.

GILLESPIE, J. H. (1991). *The Causes of Molecular Evolution.* New York: Oxford University Press.

GRADSHTEYN, I. S. & RYZHIK, I. M. (1980). *Table of Integrals, Series and Products.* San Diego: Academic Press.

HADELER, K. P. (1974). On the equilibrium states in certain selection models. *J. Math. Biol.* **1**, 51–56.

HADELER, K. P. (1981). Stable polymorphisms in a selection model with mutation. *SIAM J. Appl. Math.* **41**, 1–7.

HAIGH, J. (1978). The accumulation of deleterious genes in a population - Muller's ratchet. *Theor. Pop. Biol.* **14**, 251–267.

HEUSER, H. (1992). *Funktionalanalysis.* Stuttgart: Teubner.

HIGGS, P. G. (1994). Error thresholds and stationary mutant distributions in multi–locus diploid genetics models. *Genet. Res., Camb.* **63**, 63–78.

HOFBAUER, J. & SIGMUND, K. (1988). *The Theory of Evolution and Dynamical Systems.* Cambridge: Cambridge University Press.

HOFBAUER, J. (1985). The selection mutation equation. *J. Math. Biol.* **23**, 41–53.

HUDSON, R. R., KREITMAN, M. & AGUADÉ, M. (1987). A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**, 153–159.

HUDSON, R. R. Gene genealogies and the coalescent process. In: *Oxford Surveys in Evolutionary Biology*, volume 7, pages 1–44. Oxford University Press, (1990).

JONES, B. L., ENNS, R. H. & RANGNEKAR, S. S. (1976). On the theory of selection of coupled macromolecular systems. *Bull. Math. Biol.* **38**, 15–28.

JUDD, B. H. The *white* locus in *Drosophila melanogaster.* In: *Results and Problems in Cell Differentiation. Structure and Function of Eukaryotic Chromosomes* (Henning, W., editor), volume 14, pages 81–94. Springer, (1987).

JUKES, T. H. & CANTOR, C. R. Evolution of protein molecules. In: *Mammalian Protein Metabolism*, pages 21–132. Academic Press, New York, (1969).

KAPLAN, N. L., HUDSON, R. R. & LANGLEY, C. H. (1989). The "hitchhiking effect" revisited. *Genetics* **123**, 887–899.

KARLIN, S. & TAYLOR, H. M. (1975). *A First Course in Stochastic Processes.* San Diego: Academic Press.

KARLIN, S. (1980). The number of stable equilibria for the classical one-locus multi-allele selection model. *J. Math. Biol.* **9**, 189–192.

KIMURA, M. (1983). *The Neutral Theory of Molecular Evolution.* Cambridge: Cambridge University Press.

KIMURA, M. (1987). Molecular evolutionary clock and the neutral theory. *J. Mol. Evol.* **26**, 24–33.

KINDAHL, E. C. & AQUADRO, C. F. Levels of DNA variation are correlated with rates of recombination across the third chromosome in *Drosophila melanogaster.* Genetics (in press), (1995).

KINGMAN, J. F. C. (1988). Typical polymorphisms maintained by selection at a single locus. *J. Appl. Prob.* **25**, 113–125.

KONDRACHOV, A. S. & CROW, J. F. (1991). Haploidy or diploidy: Which is better? *Nature* **351**, 314–315.

KONDRACHOV, A. S. (1982). Selection against harmful mutations in large sexual and asexual populations. *Genet. Res.* **40**, 325–332.

LANGE, B. W., LANGLEY, C. & STEPHAN, W. (1990). Molecular evolution of *Drosophila metallothionein* genes. *Genetics* **126**, 921–932.

LANGLEY, C., MONTGOMERY, E. A. & QUATTLEBAUM, W. F. (1982). Restriction map variation in the *Adh* region of *Drosophila*. *Proc. Natl. Acad. Sci. USA* **79**, 5631–5635.

LANGLEY, C. ET AL. (1988). Naturally occuring variation in the restriction map of the *Amy* region of *Drosophila melanogaster*. *Genetics* **119**, 619–629.

LANGLEY, C. H., MACDONALD, J., MIYASHITA, N. & AGUADÉ, M. (1993). Lack of correlation between interspecific divergence and intraspecific polymorphism at the *supressor of forked* region in *Drosophila melanogaster* and *Drosophila simulans*. *Proc. Natl. Acad. Sci. USA* **90**, 1800–1803.

LANGLEY, C. H. The molecular population genetics of *Drosophila*. In: *Population Biology of Genes and Molecules* (Takahata, N. & Crow, J. F., editors), pages 75–91. Baifukan, Japan, (1990).

LEIGH-BROWN, A. J. (1983). Variation of the 87*A* heat shock locus in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* **80**, 5350–5354.

LEUTHÄUSSER, I. *Physikalische und biologische Modelle der Selbstorganisation*. Dissertation, Technische Universität Braunschweig, (1987).

LEWONTIN, R. C. & HUBBY, J. L. (1966). A molecular approach to the study of genic heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics* **54**, 595–609.

MARTÍN-CAMPOS, J. M., COMERÓN, J. M., MIYASHITA, N. & AGUADÉ, M. (1992). Intraspecific and interspecific variation at the *yellow–achaete–scute* region of *Drosophila simulans* and *Drosophila melanogaster*. *Genetics* **130**, 805–816.

MAYNARD SMITH, J. & HAIGH, J. (1974). The hitch-hiking effect of a favourable gene. *Genet. Res., Camb.* **23**, 23–35.

MIYASHITA, N. & LANGLEY, C. (1988). Molecular and phenotypic variation of the *white* locus region in *Drosophila melanogaster*. *Genetics* **120**, 199–212.

MULLER, H. J. (1950). Our load of mutations. *Amer. J. Hum. Genet.* **2**, 111–176.

MULLER, H. J. (1964). The relation of recombination to mutational advance. *Mutat. Res.* **1**, 2–9.

NEI, M. & LI, W. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. (USA)* **76**, 5269–5273.

NOWAK, M. & SCHUSTER, P. (1989). Error thresholds of replication in finite populations. Mutation frequencies and the onset of Muller's ratchet. *J. Theor. Biol.* **137**, 375–395.

OHTA, T. & KIMURA, M. (1969). Linkage disequilibrium due to random genetic drift. *Genet. Res., Camb.* **13**, 47–55.

OHTA, T. & KIMURA, M. (1975). The effect of selected linked locus on heterozygosity of neutral alleles (the hitch-hiking effect). *Genet. Res., Camb.* **25**, 313–326.

OHTA, T. Extension of the neutral mutation drift hypothesis. In: *Molecular Evolution and Polymorphism*, pages 148–167. National Institute of Genetics, Mishima, Japan, (1977).

OHTA, T. (1993). An examination of the generation–time effect on molecular evolution. *Proc.*

Natl. Acad. Sci. USA **90**, 10676–10680.

REIDYS, C., FORST, C. V. & SCHUSTER, P. Replication on neutral networks. J. Math. Biol. (submitted), (1994).

RUMSCHITZKY, D. S. (1987). Spectral properties of Eigen evolution matrices. *J. Math. Biol.* **24**, 667–680.

SCHAEFFER, S. W., AQUADRO, C. F. & LANGLEY, C. H. (1988). Restriction map variation in the *Notch* region of *Drosophila melanogaster*. *Mol. Biol. Evol.* **5**, 30–40.

SMITH, B. T. ET AL. (1976). *Matrix Eigensystem Routines – EISPACK Guide*, volume 6 of *Lecture Notes in Computer Science*. New York: Springer, 2nd edition.

SOKAL, R. R. & ROHLF, F. J. (1981). *Biometry*. San Francisco: W. H. Freeman and Co.

SPIEGELMAN, S., HARUNA, I., HOLLAND, I. B., BEAUDREAU, G. & MILLS, D. R. (1965). The synthesis of a self–propagating and infectious nucleic acid with a purified enzyme. *Proc. Natl. Acad. Sci. USA* **54**, 919–927.

STADLER, P. F., SCHNABL, W., FORST, C. V. & SCHUSTER, P. Dynamics of small autocatalytic reaction networks. II: Replication, mutation and catalysis. Bull. Math. Biol. (submitted), (1994).

STEPHAN, W. & LANGLEY, C. H. (1989). Molecular genetic varation in the centromeric region of the *X* chromosome in three *Drosophila ananassae* populations. I. Contrasts between the *vermilion* and *forked* loci. *Genetics* **121**, 89–99.

STEPHAN, W. & MITCHELL, S. J. (1992). Reduced levels of DNA polymorphism and fixed between–population differences in the centromeric region of *Drosophila ananassae*. *Genetics* **132**, 1039–1045.

SWETINA, J. & SCHUSTER, P. (1982). Self–replication with errors. A model for polynucleotide replication. *Biophys. Chem.* **16**, 329–345.

TAJIMA, F. (1989*a*). The effect of change in population size on DNA polymorphism. *Genetics* **123**, 597–601.

TAJIMA, F. (1989*b*). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595.

TAKAHATA, N., SATTA, Y. & KLEIN, J. (1992). Polymorphism and balancing selection at major histocompatibility complex loci. *Genetics* **130**, 925–938.

TAKANO, T. S., KUSAKABE, S. & MUKAI, T. (1991). The genetic structure of natural populations of *Drosophila melanogaster*. XXII. Comparative study of DNA polymorphisms in northern and southern natural populations. *Genetics* **129**, 753–761.

TARAZONA, P. (1992). Error thresholds for molecular quasispecies as phase transitions: From simple landscapes to spin–glass models. *Phys. Rev. A* **45(8)**, 6038–6050.

THOMPSON, C. J. & MCBRIDE, J. L. (1974). On Eigen's theory of the self-organization of matter and the evolution of biological macromolecules. *Math. Biosci.* **21**, 127–142.

WAGNER, G. P. & KRALL, P. (1993). What is the difference between models of error thresholds and Muller's ratchet. *J. Math. Biol.* **32**, 33–44.

WATSON, J. D. & CRICK, F. H. C. (1953). Molecular structure of nucleic acid. A structure of deoxyribose nucleic acid. *Nature* **171**, 737.

WRIGHT, S. (1955). Classification of the factors of evolution. *Cold Spring Harbor Symp. Quant. Biol.* **20**, 16–24.

Index

Selbständigkeitserklärung

Ich erkläre, daß ich die vorliegende Arbeit ("Processes Determining Genetic Variability: Mutations in Sequence Space and Hitchhiking") selbständig und nur unter Verwendung der angegebenen Hilfsmittel und Literatur angefertigt habe.

Jena, 11. Januar 1995

Lebenslauf

Name	Thomas Wiehe
Geburtsdatum	16.10.1961
Geburtsort	Coburg
Familienstand	Ledig
1968-1972	Grundschule in Coburg
1972-1981	Gymnasium Casimirianum Coburg
1981	Abitur
1982-1983	Studium der Philosophie, griechischen Philologie und Mathematik an der Universität Würzburg
1983-1989	Studium der Mathematik und Philosophie an der Universität Erlangen
1989	Diplom in Mathematik
1990	Studium des Italienischen an der Universität Siena
1990-1992	Beginn des Promotionsstudiums in Populationsgenetik und Teaching Assistant am Department of Mathematics der University of Maryland
1993-1994	Fortführung der Promotion am Institut für molekulare Biotechnologie in Jena
....	Promotion an der Friedrich Schiller Universität Jena Titel der Dissertation: *Processes Determining Genetic Variability:* *Mutations in Sequence Space and Hitchhiking*