Four-state quantum chain as a model of sequence evolution

JOACHIM HERMISSON[1,2], HOLGER WAGNER[3] AND MICHAEL BAAKE[1]

[1]Institut für Theoretische Physik, Universität Tübingen,
Auf der Morgenstelle 14, 72076 Tübingen, Germany
[2]Institut für Theorie der Kondensierten Materie,
Universität Karlsruhe, 76128 Karlsruhe, Germany
[3]Max-Planck-Institut für Biophysikalische Chemie,
Am Faßberg 11, 37077 Göttingen, Germany

February 14, 2000

Abstract

A variety of selection-mutation models for DNA (or RNA) sequences, well known in molecular evolution, can be translated into a model of coupled Ising quantum chains. This correspondence is used to investigate the genetic variability and error threshold behaviour in dependence of possible fitness landscapes and mutation mechanisms. Explicit analytic and numeric results are presented for evolution dynamics and equilibrium behaviour in a number of accessible situations, such as quadratic fitness landscapes and the Kimura 2 parameter mutation scheme.

1 Introduction

One prominent phenomenon in the theory of molecular evolution that has also attracted considerable attention in statistical physics is the so-called *error threshold*. It describes the breakdown of genetic order in mutation-selection models for mutation rates surpassing a certain critical value. The prototype model for the description of the error threshold is Eigen's quasispecies model in sequence space [7, 8] (which is effectively equivalent to a coupled mutation-selection model in population genetics, cf [5]), originally designed for the description of prebiotic RNA evolution. One important step towards an understanding of the threshold phenomenon has been its identification with an equilibrium phase transition in physics by the translation of a time-discrete version of the quasispecies model into the transfer matrix of an anisotropic two-dimensional Ising model [15]. This equivalence was further exploited to study various aspects of the error threshold with methods from statistical physics [16, 24, 11, 10, 18]. It turns out, however, that the anisotropy of that model is not so easy to handle and the analysis of the relevant biological quantities (which correspond to certain surface properties of the Ising model) remains an involved problem. Due to the complications of the model, almost all results obtained so far are approximate or numerical. The only exact result for the *sharply peaked landscape* [12] has been worked out via a different analogy to a model of directed polymers, using the specific properties of that very special fitness landscape.

An alternative approach to the analysis of mutation-selection models and the error threshold which avoids some of the problems of the anisotropic Ising model has been brought up in [2, 27]. Here, the starting point on the biological side is a slightly changed model which describes the evolution of a population with overlapping generations in continuous time. It turns out that, after a reformulation in tensor products, this model is equivalent to the Hamiltonian of an Ising quantum chain. Thereby, the change to continuous time in the biological description corresponds to the anisotropic limit that connects the two-dimensional Ising model and the quantum chain in physics (cf. [14]). The quantum chain model is technically easier to handle, and exact results for two non-trivial fitness landscapes, namely Onsager's landscape and the quadratic fitness function, have been worked out [2, 27].

In order to set up a mutation-selection model that is tractable by analytical (or at least numerical) methods, severe simplifications of the original biological situation seem to be indispensable. Analytical approaches generally have to restrict to the treatment of infinitly large populations and/or rather simple fitness functions. Another common approximation, also used in previous studies of the quasispecies and quantum chain models, amounts for the simplified representation of genotypes as binary strings. In the context of molecular evolutionary theory, this may be thought of as representing DNA or RNA strands by sequences of *purins* and *pyrimidins*, hence with only two states per site, neglecting the fact that genetic information is really given by a four-letter alphabet. In this article, we present an extension of the two-state quantum chain model, as defined in [2], to a four-state mutation-selection model which is capable to describe the full nucleotide alphabet and the standard mutation schemes of molecular genetics. In particular, the phase diagramms are discussed in detail which are more polymorphic than for the two-state model. This shows that, for a full understanding of the error threshold behaviour in molecular evolution, investigations can not be restricted entirely to the study of two-state models.

This article has been written mainly for a physical readership in that it emphasises the equivalence of models rather than the biological implications. In order to keep the exposition readable also for theoretical biologists, however, we have avoided to dwell on technical details here. For an extended presentation of methods and techniques (with regard to the two-state model), we refer to [26, 27].

In the following section, we start with a presentation of the biological foundations of our model. Only thereafter, we will introduce the quantum chain model in Section 3. In Section 4, analytical and numerical results are presented for a number of specific four-state models with permutation invariant fitness landscapes. Also the properties of finite sequences and the evolution dynamics will be studied. We close with a summary of our results and a discussion of open problems in Section 5.

2 Biological foundations

Genetic information is coded in DNA (and RNA) molecules. These are heteropolymers of four units (nucleotides) which differ in a specific base. The essential aspect of a DNA sequence is captured in a string over a four-letter alphabet

$$\boldsymbol{\sigma} \in V \equiv V_1 \times V_2 \times \cdots \times V_N \; ; \quad V_i = \{A, C, G, T\} \tag{1}$$

where each letter represents a particular base: A and G for adenine and guanine (the purins), C and T for cytosine and thymine (the pyrimidins). In RNA sequences, T is replaced by U for uracil. We will therefore treat the 4^N different sequences of a fixed, finite length N as our genotypes (which may be thought of as coding for something, such as a virus or an enzyme). Disregarding environmental effects, we may identify a collection of genotypes with a *population* of haploid 'individuals'. Evolution then describes the change of the population composition in time.

A standard model for the evolution of an infinite, asexually reproducing population under the basic forces of mutation and selection which works in continuous time is given by the following system of non-linear differential equations [5]

$$\dot{p}_{\boldsymbol{\sigma}}(t) = \big(r_{\boldsymbol{\sigma}} - \bar{r}(t)\big) p_{\boldsymbol{\sigma}}(t) + \sum_{\boldsymbol{\sigma}'} m_{\boldsymbol{\sigma}\boldsymbol{\sigma}'} p_{\boldsymbol{\sigma}'}(t) \ . \tag{2}$$

Here, $p_{\boldsymbol{\sigma}}(t)$ denotes the relative frequency of genotype $\boldsymbol{\sigma}$ at time t with corresponding Malthusian fitness (replication rate minus death rate) $r_{\boldsymbol{\sigma}}$, and

$$\bar{r}(t) = \sum_{\boldsymbol{\sigma}} r_{\boldsymbol{\sigma}} p_{\boldsymbol{\sigma}}(t) \tag{3}$$

is the *mean fitness* of the population. It is the origin of the non-linearity in (2). Finally, $m_{\boldsymbol{\sigma}\boldsymbol{\sigma}'}$ is the (time independent) rate at which $\boldsymbol{\sigma}'$ mutates to $\boldsymbol{\sigma}$. This framework has originally been defined in classical population genetics [5]. In the sequence space context, it has been introduced in [1] and has been called the *paramuse* (*pa*rallel *mu*tation-*se*lection) model, since it assumes mutation and selection to act independently and in parallel at each instant of time. The model ignores recombination and genetic drift due to finite population size. Both assumptions can be considered as fairly reasonable at least in the context of the evolution of viruses or bacteria where populations can be huge and recombination is absent, or the nucleotides are tightly linked. In the following subsections, the basic processes of mutation and selection shall be described in some detail.

2.1 Mutation

We take mutation as a point process acting independently on all sites, ignoring more complicated mechanisms, such as insertions or deletions. Molecular mutation rates shall be chosen according to the following scheme, known as the *Kimura 3 ST model* in molecular phylogeny [17, 23]:

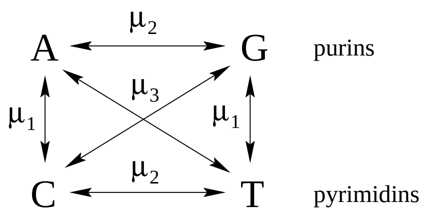

Figure 1: Molecular mutation scheme according to the Kimura 3 ST model.

Within this general setup, a number of simpler models is contained, which treat mutation at different levels of sophistication. In the simplest approach, the mutation rates between all four nucleotides are assumed to be equal ($\mu_1 = \mu_2 = \mu_3$). This is the so-called *Jukes-Cantor mutation scheme.* While this simple frame already seems to be sufficient for a number of applications, measurements reveal that there are indeed pronounced differences in the mutation rates that should be accounted for in more realistic models. In particular, the *transitions* between the two purins (A,G) and the two pyrimidins (C,T) are much more frequent than the purin–pyrimidin mutations which are called *transversions.* This may range up to relative differences of $\mu_1 \approx \mu_3 \simeq \mu_2/2$ in the nucleus and $\mu_1 \approx \mu_3 \simeq \mu_2/40$ in mitochondrial DNA [17]. A mutation scheme with $\mu_2 > \mu_1 = \mu_3$ is known as the *Kimura 2 parameter model.* The full *Kimura 3 ST* scheme, finally, also accounts for the small difference between μ_1 and μ_3, such that $\mu_2 > \mu_1 > \mu_3$.

Implementing this mutation model into the evolution equation (2), we obtain the following mutation rates between genotypes ($i \in \{1,2,3\}$)

$$m_{\boldsymbol{\sigma}\boldsymbol{\sigma}'} = \begin{cases} \mu_i, & d_i(\boldsymbol{\sigma},\boldsymbol{\sigma}') = d_{\boldsymbol{\sigma}\boldsymbol{\sigma}'} = 1 \\ -N\sum_i \mu_i, & \boldsymbol{\sigma} = \boldsymbol{\sigma}' \\ 0, & d_{\boldsymbol{\sigma}\boldsymbol{\sigma}'} > 1 \end{cases} . \tag{4}$$

Here,

$$\begin{aligned} d_1(\boldsymbol{\sigma},\boldsymbol{\sigma}') &= \#_{A\rightleftarrows C}(\boldsymbol{\sigma},\boldsymbol{\sigma}') + \#_{G\rightleftarrows T}(\boldsymbol{\sigma},\boldsymbol{\sigma}') \\ d_2(\boldsymbol{\sigma},\boldsymbol{\sigma}') &= \#_{A\rightleftarrows G}(\boldsymbol{\sigma},\boldsymbol{\sigma}') + \#_{C\rightleftarrows T}(\boldsymbol{\sigma},\boldsymbol{\sigma}') \\ d_3(\boldsymbol{\sigma},\boldsymbol{\sigma}') &= \#_{A\rightleftarrows T}(\boldsymbol{\sigma},\boldsymbol{\sigma}') + \#_{C\rightleftarrows G}(\boldsymbol{\sigma},\boldsymbol{\sigma}') \end{aligned} \tag{5}$$

are restricted Hamming distances between $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}'$. In (5), $\#_{X\rightleftarrows Y}(\boldsymbol{\sigma},\boldsymbol{\sigma}')$ counts the positions at which X and Y are exchanged in $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}'$. Finally,

$$d_{\boldsymbol{\sigma}\boldsymbol{\sigma}'} = d_1(\boldsymbol{\sigma},\boldsymbol{\sigma}') + d_2(\boldsymbol{\sigma},\boldsymbol{\sigma}') + d_3(\boldsymbol{\sigma},\boldsymbol{\sigma}') \tag{6}$$

is the total Hamming distance. Note that the choice of the diagonal term $m_{\boldsymbol{\sigma}\boldsymbol{\sigma}}$ in (4) just accounts for probability conservation ($\sum_{\boldsymbol{\sigma}} \dot{p}_{\boldsymbol{\sigma}} = 0$) in the mutation part of the evolution equation (2).

2.2 Selection and fitness landscape

Whereas the mutational part of the dynamics is fairly well understood at least on the microscopic (molecular) level, the relation of genotype and fitness, which defines the respective selective success, is notoriously complex. Following the standard notion in molecular evolution, we define the *fitness function* (or *fitness landscape*)

$$f : \boldsymbol{\sigma} \mapsto r_{\boldsymbol{\sigma}} \tag{7}$$

as a mapping from the configuration space $V = \{A, C, G, T\}^N$ into the real numbers, assigning a reproduction rate (Malthusian fitness value) $r_{\boldsymbol{\sigma}}$ to each genotype. Implicitly, the fitness function incorporates all the complicated interactions between the sites. These interactions are typically long-ranged (since RNA strands or proteins fold in three dimensions), highly correlated, and give rise to rather rugged landscapes. Especially in the context of RNA evolution, the construction and characterization of fitness landscapes has motivated numerous studies, see e.g. [22] for a review.

Below we will show how the evolution equation (2), with an arbitrary choice of the fitness function, can be adapted to the methods from statistical physics by a reformulation in a quantum chain framework. As an application, we then present a study (including analytical and numerical results) for specific examples from the class of permutation invariant fitness functions. Here, due to equivalence of all sites, the fitness of a given genotype is solely a function of its restricted Hamming distances from the so called *wildtype* sequence with optimal fitness which we choose as the reference genotype. This particularly simple class of fitness landscapes is widely used, as a canonical first approximation, especially in *multilocus theory*. Also in the context of sequence space evolution, fitness functions of this type have been used in a number of studies on the two-state model [20, 16, 24, 2, 27]. To implement the approach in our four-state model, we fix an arbitrary sequence, denoted by $\boldsymbol{\sigma}_{++}$, as the wildtype. We will only consider directional selection here towards a unique genotype with optimal fitness. The fitness of any other sequence is then determined by the restricted Hamming distances d_i relative to $\boldsymbol{\sigma}_{++}$. Permutation invariance with respect to the position in the sequence thus leads to a drastic reduction of dimensions. For the four-state model, the effective configuration space

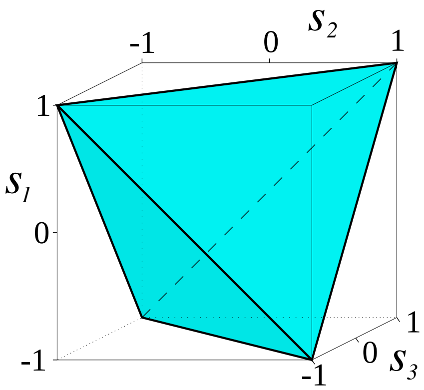

Figure 2: Permutation invariant configuration space of the four-state model in surplus coordinates.

forms a tetrahedron in 3d (see Fig. 2) and is conveniently represented in Cartesian coordinates which we shall call (following [2]) the *surplus components*:

$$\begin{aligned} s_1(\boldsymbol{\sigma}) &= 1 - \frac{2}{N}\Big(d_1(\boldsymbol{\sigma},\boldsymbol{\sigma}_{++}) + d_3(\boldsymbol{\sigma},\boldsymbol{\sigma}_{++})\Big) ; \\ s_2(\boldsymbol{\sigma}) &= 1 - \frac{2}{N}\Big(d_2(\boldsymbol{\sigma},\boldsymbol{\sigma}_{++}) + d_3(\boldsymbol{\sigma},\boldsymbol{\sigma}_{++})\Big) ; \\ s_3(\boldsymbol{\sigma}) &= 1 - \frac{2}{N}\Big(d_1(\boldsymbol{\sigma},\boldsymbol{\sigma}_{++}) + d_2(\boldsymbol{\sigma},\boldsymbol{\sigma}_{++})\Big) . \end{aligned} \tag{8}$$

With this choice, any unstructured random sequence has coordinates $s_i \equiv 0$ (with probability 1 in the limit $N \to \infty$). Any positive value of a surplus component, on the other hand, signals a non-trivial overlap of the sequence with the wildtype $\boldsymbol{\sigma}_{++}$. In particular, s_1 measures the surplus of sites with purins or pyrimidins as given in $\boldsymbol{\sigma}_{++}$ over the purin–pyrimidin mutated sites.

Within this frame, a natural class of permutation invariant fitness functions is

$$f : \boldsymbol{\sigma} \mapsto r_{\boldsymbol{\sigma}} = N \sum_{i=1}^{3} \left[\alpha_i s_i(\boldsymbol{\sigma}) + \frac{\gamma_i}{2} s_i^2(\boldsymbol{\sigma})\right] \tag{9}$$

which includes the following special cases

- Setting $\alpha_i > 0$ and $\gamma_i = 0$, we obtain the purely additive *Fujiyama landscape* without genetic interactions. Here, every mutation relative to the wildtype has a fixed deleterious effect, independent of any other mutation that may be present in the sequence. The additive landscape is a canonical zeroth-order approximation, ignoring any kind of genetic interactions. In the context of sequence evolution, this fitness function has been discussed e.g. in [20, 2].

- With the choice $\alpha_i \geq -\gamma_i > 0$, the model corresponds to a concave quadratic fitness function (with directional selection) as it is frequently met in multilocus theory. Due to the gene interactions, existing mutations tend to aggravate further ones, which is called *positive epistasis*.

- For $\alpha_i \geq 0$ and $\gamma_i > 0$, we finally obtain a convex fitness function for directional selection with long-range gene interactions and *negative epistasis* (existing mutations tend to alleviate further ones). Since we want to have $\boldsymbol{\sigma}_{++}$ as unique wildtype sequence and a fitness function which is monotonous in the surplus components, we restrict f to the octant $s_i \geq 0$ and (smoothly) truncate the fitness function by introduction of a step function $\Theta(s_i)$ whenever frequencies of genotypes with $s_i < 0$ are non-zero:

$$\tilde{f} : \boldsymbol{\sigma} \mapsto r_{\boldsymbol{\sigma}} = N \sum_{i=1}^{3} \left[\left(\alpha_i s_i(\boldsymbol{\sigma}) + \frac{\gamma_i}{2} s_i^2(\boldsymbol{\sigma}) \right) \Theta(s_i) \right] . \tag{10}$$

The variables α_i and γ_i may further be used to distinguish between the effects of the different types of mutations (as defined in Fig 1) on the fitness. In this article, we will present explicit results for the two following cases:

1. For the simplest choice, $\alpha_1 = \alpha_2 = \alpha_3$ and $\gamma_1 = \gamma_2 = \gamma_3$, any mutation away from the wildtype has the same effect. Together with the Jukes-Cantor mutation scheme, symmetry here leads to equal values of the surplus components in the mutation–selection equilibrium. The model may thus also be thought of as a two-state model, where any site is only regarded as occupied either with a *wildtype* or with a *mutant* nucleotide. In contrast to the simple two-state model of [2], however, there is an effectively asymmetric mutation rate between wildtype and mutant in the case considered here.

2. In a more refined model, we distinguish between transitions and transversions. In the mutational part, this is done by applying the Kimura 2 parameter mutation scheme. In the fitness function, we take into account that the deleterious effects of the transversions often dominate over those of the transitions: $\alpha_1 > \alpha_{2,3}$ and/or $\gamma_1 > \gamma_{2,3}$.

3 Quantum chain model

3.1 Symmetries

Since mutation is a random process that is independent of the fitness values of the genotypes involved, the molecular mutation scheme consequently makes no reference to fitness concepts like the *wildtype*. Biological observables measurable from sequence data, such as the surplus components (8), and also the fitness functions as defined in (9) or (10), on the other hand, are defined relative to the wildtype sequence. In order to set up these concepts in a common framework, it is convenient to reformulate also the mutational part of the evolution equation in coordinates relative to the wildtype. This may always be done due to certain symmetries inherent in the mutation scheme of Fig. 1.

The basic symmetry of the mutation scheme, if all three mutation rates μ_1, μ_2, μ_3 are pairwise different, is $C_2 \times C_2$ (Klein's 4-group), generated by two involutions. If we write the operations in standard permutation notation, we can take as generators the transformations

$$\begin{pmatrix} A & C & G & T \\ C & A & T & G \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} A & C & G & T \\ G & T & A & C \end{pmatrix} , \tag{11}$$

both being the product of two transpositions. This symmetry may now be exploited for a redefinition of the mutation scheme in wildtype coordinates. To this end, we fix, for every site of the wildtype sequence, the element of the 4-group (in the above representation) with the letter of the wildtype nucleotide in the first position (e.g. the string (T, G, C, A) for wildtype nuceotide T). An alternative representation of the configuration space in wildtype coordinates as

$$\boldsymbol{\sigma} \in V^{\pm} \equiv V_1^{\pm} \times V_2^{\pm} \times \cdots \times V_N^{\pm} \; ; \quad V_i^{\pm} = \{++, -+, +-, --\} \tag{12}$$

is now given by the mapping, on each site, of the string of labels $(++,-+,+-,--)$ to the symmetry element of 4-group defined above. With this notation, the three types of mutations included in the Kimura 3 ST scheme simply switch the signs of the labels: $\pm\pm \to \mp\pm$ at rate μ_1, $\pm\pm \to \pm\mp$ at rate μ_2, and $\pm\pm \to \mp\mp$ at rate μ_3.

Higher symmetries of the mutation model are obtained if mutation rates are equal. For the Kimura 2 parameter scheme, $\mu_1 = \mu_3 \neq \mu_2$, the operation

$$A \to C \to G \to T \to A \;=\; \begin{pmatrix} A & C & G & T \\ C & G & T & A \end{pmatrix} \tag{13}$$

is also a symmetry and generates a cyclic group C_4. Together with the previous $C_2 \times C_2$, this generates a dihedral group, D_4, with 8 elements. Finally, if $\mu_1 = \mu_2 = \mu_3$, we additionally get the simple transposition $A \leftrightarrow C$ and have the full permutation group S_4 as symmetry. Note that S_4, which corresponds to the full tetrahedral group with 24 elements, is also the symmetry group of the configuration space of permutation invariant configurations visualized in Fig. 2. The *global* symmetry (with the same transformation acting at each site simultaneously) of our class of mutation-selection models with fitness functions according to (9) is therefore always a subgroup of S_4. In particular, the symmetric fitness model with $\alpha_1 = \alpha_2 = \alpha_3$, $\gamma_1 = \gamma_2 = \gamma_3$, and Jukes-Cantor mutation scheme possesses C_{3v} symmetry, or the full tetrahedral symmetry if the linear part in the fitness function vanishes ($\alpha_i = 0$). The transition-transversion model finally, with $\alpha_1 > \alpha_2 = \alpha_3$, or $\gamma_1 > \gamma_2 = \gamma_3$, and Kimura 2 parameter mutation has simple C_2 symmetry, or D_4 symmetry if $\alpha_i \equiv 0$. In the latter case, the combination of $\gamma_2 = \gamma_3$ with $\mu_1 = \mu_3$ is necessary, not a misprint. Other combinations with global D_4 symmetry are $(\gamma_1 = \gamma_3; \mu_2 = \mu_3)$ and $(\gamma_1 = \gamma_2; \mu_1 = \mu_2)$.

3.2 Construction

With the above preparations, we may now follow the lines of [2, 27] where the two-state model is treated.

In a first step, we represent the 4^N-dimensional vector space in which we describe the genotype frequencies as the N-fold tensor product space $W = \otimes_{j=1}^N W_j$. Hereby, the configuration space $V^\pm$ is canonically embedded in W by the mapping of the elements of $V_i^\pm$ onto the basis vectors $\{e_j^{++}, e_j^{-+}, e_j^{+-}, e_j^{--}\}$ of $W_j \simeq \mathbb{R}^4$. Since the nonlinear part in the differential equations (2) only amounts to normalization of the frequencies, a transformation to so-called *absolute frequencies* [25, 2]

$$z_{\boldsymbol{\sigma}}(t) = p_{\boldsymbol{\sigma}}(t) \exp\Big(\sum_{\boldsymbol{\sigma}'} r_{\boldsymbol{\sigma}'} \int_0^t p_{\boldsymbol{\sigma}'}(\tau)\, d\tau\Big) \tag{14}$$

then reduces the system to the linear equation

$$\dot{z}_{\boldsymbol{\sigma}}(t) = \big(\mathcal{M} + \mathcal{R}\big) z_{\boldsymbol{\sigma}}(t) \tag{15}$$

where the mutation and reproduction matrices, $\mathcal{M} = (m_{\boldsymbol{\sigma}\boldsymbol{\sigma}'})$ and $\mathcal{R} = \mathrm{diag}(r_{\boldsymbol{\sigma}})$, may now be conveniently represented in the frequency space W. Defining

$$\sigma_j^{(\alpha,\beta)} := \big(\otimes^{j-1} \mathbb{1}_4\big) \otimes \big(\sigma^\alpha \otimes \sigma^\beta\big) \otimes \big(\otimes^{N-j-1} \mathbb{1}_4\big) \tag{16}$$

where σ^α, $\alpha \in \{0, x, z\}$, are the real Pauli matrices and $\sigma^0 \equiv \mathbb{1}_2$, we find

$$\mathcal{M} = \sum_{j=1}^N \Big[\mu_1 \sigma_j^{(x,0)} + \mu_2 \sigma_j^{(0,x)} + \mu_3 \sigma_j^{(x,x)} - (\mu_1 + \mu_2 + \mu_3)\mathbb{1}\Big] \tag{17}$$

for the mutation matrix. The reproduction matrix $\mathcal{R}$ is, for a general fitness landscape, an element of the algebra generated by $\sigma_j^{(z,0)}$ and $\sigma_j^{(0,z)}$, $1 \leq j \leq N$,

$$\mathcal{R} = r_0 \mathbb{1} + \sum_{k,\ell=1}^{N} \sum_{[j_1 \cdots j_k]} \sum_{[j_1 \cdots j_\ell]} \varepsilon_{[j_1 \cdots j_k],[j_1 \cdots j_\ell]} \prod_{m=1}^{k} \sigma_{j_m}^{(z,0)} \prod_{n=1}^{\ell} \sigma_{j_n}^{(0,z)}, \tag{18}$$

where $[j_1 \ldots j_k]$ is an ordered k-tupel in $\{1, \ldots, N\}$. Now, from a physical point of view, $\mathcal{H} = \mathcal{M} + \mathcal{R}$ is (up to a global minus sign) the Hamiltonian of two coupled Ising quantum chains in a tunable transverse magnetic field (the mutation) and general spin-interactions within the chains.

Translated to our quantum chain model, the fitness function of the permutation invariant landscape defined in (9) results in a (longitudinal) magnetic field and a mean field spin-interaction. We find $\mathcal{R} = \mathcal{R}_\alpha + \mathcal{R}_\gamma$, where

$$\mathcal{R}_\alpha = \sum_{j=1}^{N} \left[\alpha_1 \sigma_j^{(z,0)} + \alpha_2 \sigma_j^{(0,z)} + \alpha_3 \sigma_j^{(z,z)} \right] \tag{19}$$

and

$$\mathcal{R}_\gamma = \frac{1}{2N} \sum_{j,k=1}^{N} \left[\gamma_1 \sigma_j^{(z,0)} \sigma_k^{(z,0)} + \gamma_2 \sigma_j^{(0,z)} \sigma_k^{(0,z)} + \gamma_3 \sigma_j^{(z,z)} \sigma_k^{(z,z)} \right] \tag{20}$$

Let us stress that, in contrast to most physical applications, the mean field model is a much more natural approach in the biological context where interactions are typically long-range. So, it is a legitimate model here, not an inevitable approximation.

3.3 Biological and physical observables

In this subsection, we relate the quantities of biological interest, mean and variance of the surplus components and the fitness, to the physical observables. In what follows, we assume the occuring limits to exist.

Genotype composition According to (15), the Hamiltonian of the quantum chain determines the time evolution of our population of genotypes in an environment that does not constrain the population size. For any genotype-independent regulation of the population size, the relative genotype frequencies are found by *statistical* normalization. We therefore define the vector of the genotype composition $|\boldsymbol{p}(t)\rangle$ and the equilibrium composition $|0\rangle$ as

$$|\boldsymbol{p}(t)\rangle = \frac{\exp(t\mathcal{H})|\boldsymbol{p}_0\rangle}{\langle\Omega|\exp(t\mathcal{H})|\boldsymbol{p}_0\rangle} \quad ; \quad |0\rangle := \lim_{t\to\infty} |\boldsymbol{p}(t)\rangle \tag{21}$$

where $|\boldsymbol{p}_0\rangle$ is the initial composition and $4^{-N}|\Omega\rangle$ is the equidistribution of genotypes. Note that the *equilibrium composition* of the genotype population just corresponds to the *ground state* of the quantum chain on the physical side (with a different 'biological' normalization $\langle\Omega|0\rangle = 1$).

Fitness The *density of the mean fitness* (or mean fitness per site) of the population is given by the expression

$$w(t) := N^{-1}\bar{r}(t) = N^{-1}\langle\Omega|\mathcal{R}|\boldsymbol{p}(t)\rangle \ . \tag{22}$$

Since

$$w := \lim_{t\to\infty} w(t) = N^{-1}\langle\Omega|\mathcal{R}|0\rangle = N^{-1}\frac{\langle 0|\mathcal{H}|0\rangle}{\langle 0|0\rangle} \tag{23}$$

the *equilibrium* mean fitness (per site) is just given by the (unique) largest eigenvalue of $\mathcal{H}$, corresponding to $|0\rangle$. For an unconstrained population, w also determines the growth rate in the long-time limit. In the physical picture, $(-w)$ is obviously just the *ground state energy* (per spin).

Using $\mathcal{M}|\Omega\rangle = 0$, we derive for the time evolution of the mean fitness

$$\dot{w}(t) = V_r(t) + N^{-1}\langle\Omega|[\mathcal{R},\mathcal{M}]|\boldsymbol{p}(t)\rangle \tag{24}$$

where $V_r(t)$ is the *variance of fitness* (per site),

$$V_r(t) = \frac{1}{N}\left(\langle\Omega|\mathcal{R}^2|\boldsymbol{p}(t)\rangle - \langle\Omega|\mathcal{R}|\boldsymbol{p}(t)\rangle^2\right) \ . \tag{25}$$

In the absence of mutation, (24) is of course just a special case of Fisher's "Fundamental Theorem of Natural Selection" [9] which states that the rate of increase in fitness is equal to the genetic variance in fitness. For the mutation-selection models considered here, the relation has the following intuitive interpretation: The change in mean fitness is driven by two independent forces. The first one stems from the change of genotype frequencies due to selection and is proportional to the variance of fitness values present in the population. Since variances are positive, it always tends to increase fitness. The second term on the right hand side of (24) typically decreases fitness. It measures the population mean of the change in fitness at time t due to the action of mutation. In mutation-selection equilibrium, both terms balance, and the entire residual variance is due to mutation.

Surplus Another quantity that characterizes the genetic order of the population, as it may be measured from sequence data, is the *mean surplus*. We define, following and generalizing [2],

$$u_i(t) = \sum_{\boldsymbol{\sigma}} s_i(\boldsymbol{\sigma}) p_{\boldsymbol{\sigma}}(t) \quad ; \quad u_i = \lim_{t\to\infty} u_i(t) \ . \tag{26}$$

In particular,

$$\#_m(t) := \frac{1}{4}\big(3 - (u_1(t) + u_2(t) + u_3(t))\big) \tag{27}$$

measures the mean number of mutations per site relative to the wildtype while

$$\#_{tr}(t) := \frac{1}{2}\big(1 - u_1(t)\big) \tag{28}$$

denotes the mean number of transversions alone. As a *biological order parameter*, the mean surplus plays a similar rôle as the physical magnetization. However, as already noted in [3], both quantities are quite distinct and in many cases not even easily related. In the language of the quantum chain, the equilibrium mean surplus may be derived as

$$u_1 = \frac{\langle\Omega|\sum_i \sigma_i^{(z,0)}|0\rangle}{N} \quad ; \quad u_2 = \frac{\langle\Omega|\sum_i \sigma_i^{(0,z)}|0\rangle}{N} \quad ; \quad u_3 = \frac{\langle\Omega|\sum_i \sigma_i^{(z,z)}|0\rangle}{N} \ , \tag{29}$$

whereas the three-component magnetization is defined as the ground state expectation value

$$m_1 = \frac{\langle 0|\sum_i \sigma_i^{(z,0)}|0\rangle}{N\langle 0|0\rangle} \quad ; \quad m_2 = \frac{\langle 0|\sum_i \sigma_i^{(0,z)}|0\rangle}{N\langle 0|0\rangle} \quad ; \quad m_3 = \frac{\langle 0|\sum_i \sigma_i^{(z,z)}|0\rangle}{N\langle 0|0\rangle} \ . \tag{30}$$

As we will show below, magnetization and surplus can show rather different behaviour especially near phase transitions. The biological and physical phase diagrams, however, coincide if phase transitions (or error thresholds) are defined as nonanalyticity points of the ground state energy (or mean fitness) w in the thermodynamic limit (cf. the discussion in Section 5).

4 Results

4.1 Fujiyama model

As in the two-letter case [2], the quantum chain model decomposes into non-interacting one-site Hamiltonians for the additive landscape. The mean fitness and its variance are linear functions in the surplus components. In particular, we obtain from (24)

$$V_r(t) = \dot{w}(t) + 2\big((\mu_1+\mu_3)\alpha_1 u_1(t) + (\mu_2+\mu_3)\alpha_2 u_2(t) + (\mu_1+\mu_2)\alpha_3 u_3(t)\big) \; . \quad (31)$$

For Jukes-Cantor mutation, $\mu_1 = \mu_2 = \mu_3 \equiv \mu$, this reduces to

$$V_r(t) = \left(4\mu + \frac{\mathrm{d}}{\mathrm{d}t}\right) w(t) \quad (32)$$

and V_r is proportional to the mean fitness in the mutation–selection equilibrium. Exact results are easily found from the solution of the four-dimensional eigenvalue problem of the one-site Hamiltonian. We only give the expression for the mean fitness in the symmetric case, $\alpha_1 = \alpha_2 = \alpha_3 \equiv \alpha$ with Jukes-Cantor mutation scheme ($\mu_1 = \mu_2 = \mu_3 \equiv \mu$):

$$w(t) = \frac{\exp[2t(\alpha+\mu)]\cosh[2tQ]\,(\alpha - 2\mu + 2Q\tanh[2tQ]) - \alpha - 4\mu}{1 + \exp[2t(\alpha+\mu)]\cosh[2tQ]} \quad (33)$$

where

$$Q = \sqrt{\mu^2 + \alpha^2 - \alpha\mu} \quad (34)$$

and the equidistribution of genotypes is chosen as starting configuration.

Means and variances of the fitness and the surplus in mutation–selection balance are shown in Fig. 6 below. A plot of the time evolution of fitness is given in Fig. 8. There is clearly no phase transition (resp. no *error threshold* behaviour) for the additive Fujiyama landscape, as expected in view of the complete absence of interactions (resp. epistasis).

4.2 Quadratic fitness model: Equilibrium results

In contrast to the additive case, no simple relation between surplus and fitness is known in the case of the quadratic landscape as long as t or N are kept finite. However, due to the permutation invariance of the Hamiltonian, the individual fitness–surplus relation (9) is recovered in the thermodynamic limit for the corresponding mean values of the equilibrium population. We obtain in analogy to [3]:

$$w = \lim_{t\to\infty} w(t) = \sum_{i=1}^{3}\left(\alpha_i u_i + \frac{\gamma_i}{2}u_i^2\right) \quad (35)$$

and, from (24), for the equilibrium variance of fitness per site

$$V_r = \lim_{t\to\infty} V_r(t) = 2(\mu_1+\mu_3)\left(\alpha_1 u_1 + \gamma_1 u_1^2\right) + \\ 2(\mu_2+\mu_3)\left(\alpha_2 u_2 + \gamma_2 u_2^2\right) + 2(\mu_1+\mu_2)\left(\alpha_3 u_3 + \gamma_3 u_3^2\right) \ . \quad (36)$$

The key to the solution in the thermodynamic limit is now the minimum principle of the physical free energy which translates to a maximum principle for the equilibrium mean fitness. Maximizing

$$\langle \boldsymbol{x}|\mathcal{M}+\mathcal{R}|\boldsymbol{x}\rangle - w\big(\langle \boldsymbol{x}|\boldsymbol{x}\rangle - 1\big) \quad (37)$$

with respect to w and $\boldsymbol{x}$, we obtain, taking permutation symmetry of $\boldsymbol{x}$ into account, the following variational expression for w:

$$\begin{aligned} w(\boldsymbol{\alpha},\boldsymbol{\mu},\boldsymbol{\gamma}) \ = \sup_{m_1,m_2,m_3} \Big[& \alpha_1 m_1 + \alpha_2 m_2 + \alpha_3 m_3 + \frac{\gamma_1}{2} m_1^2 + \frac{\gamma_2}{2} m_2^2 + \frac{\gamma_3}{2} m_3^2 + \\ & \frac{\mu_1}{2}\left(\sqrt{(1+m_2)^2-(m_1+m_3)^2} + \sqrt{(1-m_2)^2-(m_1-m_3)^2} - 2\right) + \\ & \frac{\mu_2}{2}\left(\sqrt{(1+m_1)^2-(m_2+m_3)^2} + \sqrt{(1-m_1)^2-(m_2-m_3)^2} - 2\right) + \\ & \frac{\mu_3}{2}\left(\sqrt{(1+m_3)^2-(m_1+m_2)^2} + \sqrt{(1-m_3)^2-(m_1-m_2)^2} - 2\right) \Big] \end{aligned} \quad (38)$$

where $m_i \in [-1,1]$ are the components of the physical magnetization. Let us stress that, from the biological point of view, the translation to the physical framework seems a necessary technical step since we do not know of any variational principle for the biological model which works directly in L^1. We now take a closer look at two special cases.

Symmetric fitness model For the symmetric *wildtype–mutant* model with $\alpha_i \equiv \alpha$, $\gamma_i \equiv \gamma$ and Jukes-Cantor mutation rate μ, all components of the order parameters are equal, $m_i \equiv m$ and $u_i \equiv u$, respectively. Here, the variational expression (38) for w leads to the following self-consistency condition for m:

$$m = \frac{1}{3}\left[1 + \frac{2(\alpha+\gamma m)-\mu}{\sqrt{(\alpha+\gamma m)^2 - \mu(\alpha+\gamma m)+\mu^2}}\right] \ . \quad (39)$$

This is a quartic equation in m and can be solved using the standard formulas. However, since the explicit solution is rather lengthly, we do not include it here, but give a qualitative discussion instead.

Obviously, the relation has a unique real solution for any α and μ whenever γ is *negative*. Like in the case of the two-state model, we thus obtain no phase transition for positive epistasis. In the following, we therefore concentrate our discussion on positive γ (or negative epistasis). Note that, for calculations in the thermodynamic limit, always the fitness function f (9), and hence the reproduction matrix $\mathcal{R}_\gamma$ (20), can be used instead of the truncated form $\tilde{f}$ (10), since the frequencies of genotypes with negative surplus vanish. For $\alpha_i \equiv 0$, this is due to spontaneous breaking of the extra $C_2 \times C_2$ symmetry of $\mathcal{H} = \mathcal{M} + \mathcal{R}_\gamma$.

In contrast to the two-state model, where a phase transition in the thermodynamic limit is only found for zero external field, it turns out that the present model has phase transitions for a whole range of the linear fitness parameter α when epistasis is negative: For $\tilde{\alpha} := \alpha/\gamma$ in the interval

$$0 \le \tilde{\alpha} < \frac{1}{3}\left(\sqrt{\frac{4}{3}} - 1\right) \simeq 0.0515668 \quad (40)$$

we find a first order phase transition of the system at

$$\tilde{\mu} := \frac{\mu}{\gamma} = \tilde{\mu}_c = \frac{2}{3} + 2\tilde{\alpha} \tag{41}$$

with a finite jump in the magnetization from m_+ to m_- where

$$m_{\pm} = \frac{1}{3}\left(1 \pm \sqrt{1 - 27\tilde{\alpha}^2 - 18\tilde{\alpha}}\right) . \tag{42}$$

From m we derive the mean fitness w using (38), from w we obtain the surplus u via (35) and, finally, the variance of the fitness $V_r = 12\mu(\alpha u + \gamma u^2)$. Looking at the surplus u, we also find a phase transition at $\tilde{\mu} = \tilde{\mu}_c$. As m, it vanishes in the disordered phase for $\alpha = 0$. Note however that, since w is continuous, due to the relation (35), also the surplus is continuous at a phase transition. In [3] it has been shown that these differences of the biological and physical order parameters arise with the change from classical to quantum mechanical probabilities (resp. the change from L^1 to L^2) in translating the biological model into the physical one. We remark that a different, discontinuous behaviour of the biological order parameter at a (physical) first order transition has been observed for the sharply peaked landscape in Eigen's quasispecies model [10]. Mean fitness and its variance, magnetization, and surplus for different values of α are shown below in Fig. 3.

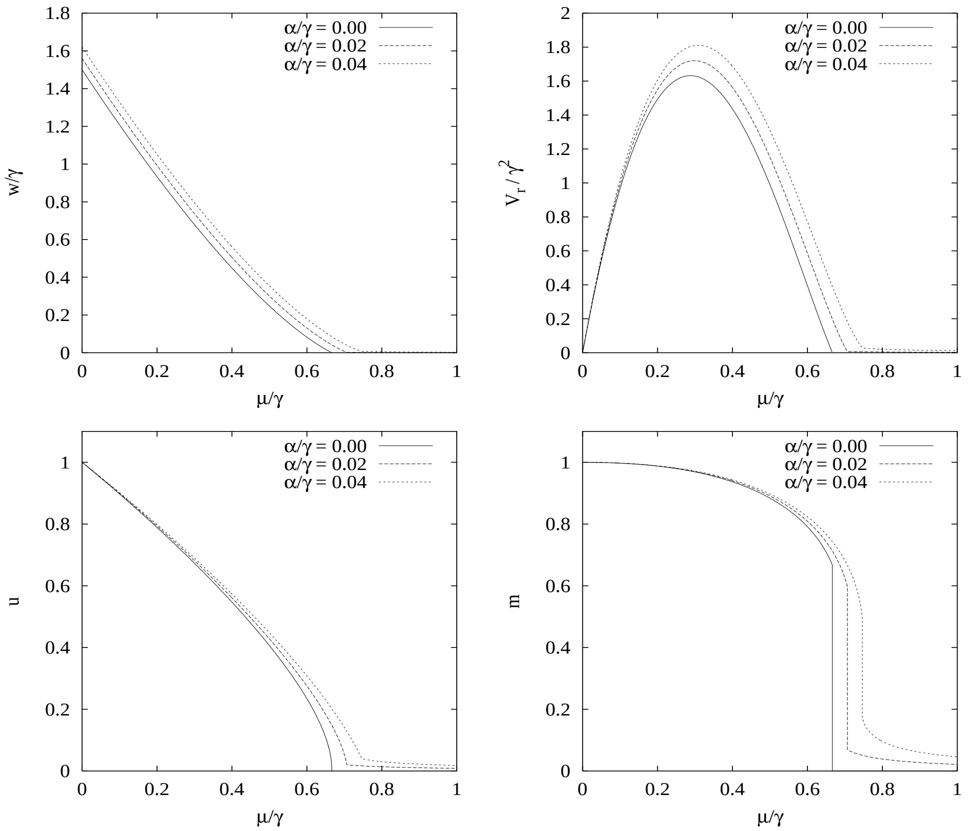

Figure 3: Mean fitness and its variance, surplus and magnetization in the symmetric fitness model for various linear parts of the fitness function in the infinite sites limit.

Transition–transversion model In our second example, we wish to distinguish mutations between like and unlike nucleotides. In a first step, we retain the symmetric fitness landscape $\gamma_1 = \gamma_2 = \gamma_3 \equiv \gamma$ (for simplicity with vanishing linear part

$\alpha = 0$), but let the relative frequencies of transitions and transversions differ by assuming the *Kimura 2 parameter* mutation scheme, $\mu_1 = \mu_3 \equiv \mu \neq \mu_2$.

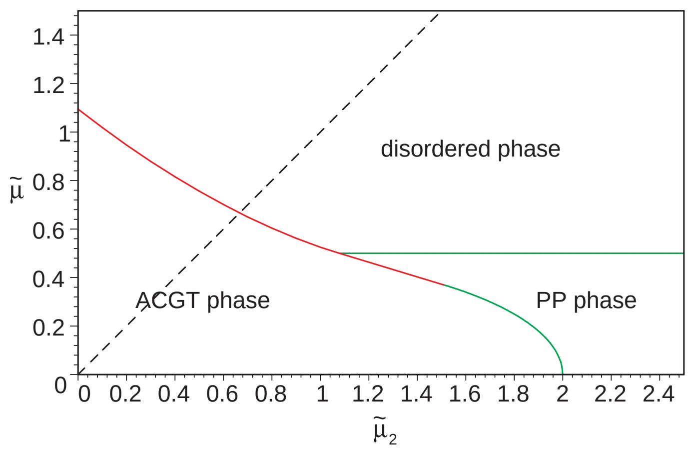

Figure 4: Phase diagram of the transition–transversion model with with symmetric fitness landscape and Kimura 2 parameter mutation scheme. Solid and dotted lines correspond to first and second order phase transitions, respectively. The dashed line indicates the Jukes-Cantor mutation scheme.

In the extended parameter space of the reduced mutation rates $\tilde{\mu} = \mu/\gamma$; $\tilde{\mu}_2 = \mu_2/\gamma$, we now obtain a phase diagram with *three* distinct phases (see Fig. 4).

- For $\tilde{\mu}$ and $\tilde{\mu}_2$ sufficiently small, all three surplus components are positive, indicating genetic order with respect to the entire 4-letter alphabet of the nucleotides: *ACGT phase*.
- If we increase the mutation rate $\tilde{\mu}_2$ for low $\tilde{\mu}$, the system crosses over to a phase which does no longer distinguish between the different kinds of purins (A,G) and pyrimidins (C,T), but is still ordered with respect to transversions. This is the limiting case described by the two-state model. We call this the *PP phase*.
- For higher mutation rates $\tilde{\mu}, \tilde{\mu}_2$, we finally enter a completely *disordered phase* with vanishing fitness and surplus.

In a second step, we now also let the mutation effects of transitions and transversions differ and assume a fitness landscape with $\gamma_2 = \gamma_3 \equiv \gamma$, but $\gamma_1 \neq \gamma$ in general. The changes in the phase diagram for increasing $\tilde{\gamma}_1 = \gamma_1/\gamma$ are shown in Fig. 5. The phase transitions between the three phases may be first or second order. In general, we obtain the following phase space structure:

- Phase transitions between the disordered and PP phase are second order and located on the line $\tilde{\mu} = \tilde{\gamma}_1/2$. This phase transition corresponds to the one also seen in the two-state model [2].
- The phase transition line between the ACGT and PP phases in general changes from first to second order with increasing mutation rate $\tilde{\mu}_2$ (see Figs. 4, 5). For the second order transitions we derive, on expanding (38) to lowest order in $m_2 = m_3$,

$$\mu = \frac{\gamma_1}{\gamma_1 + 2\gamma} \sqrt{(\gamma_1 + \mu_2)(2\gamma - \mu_2)} \; . \tag{43}$$

Numerically, we find that the first order transitions are located on a straight line up to $\tilde{\mu} = \tilde{\gamma}_1/2$ where the PP phase changes into the disordered phase. The $\tilde{\mu}_2$-interval of first-order transitions decreases for increasing $\tilde{\gamma}_1$. For $\tilde{\gamma}_1 \gtrapprox 8.45$, all phase transitions between the ACGT and PP phases are second order.

- Finally, for $\tilde{\gamma}_1 \leq 4$, there are direct first order phase transitions between the ACGT phase and the disordered phase (for $\tilde{\mu}_2$ sufficiently small). For higher values of $\tilde{\gamma}_1$, these two phases are separated by the PP phase.

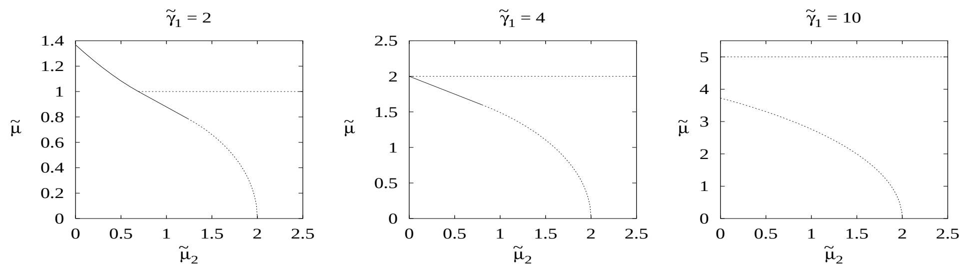

Figure 5: Phase diagrams for anisotropic fitness landscapes $\gamma_1 > \gamma_2 = \gamma_3 \equiv \gamma$ and Kimura 2 parameter mutation scheme. Solid and dotted lines correspond to first and second order phase transitions, respectively.

As for the symmetric fitness function discussed above, there are no compact analytic expressions for the fitness or the surplus in the ACGT phase. In the PP phase, however, the following values for the mean fitness and the non-zero components of the mean surplus and the magnetization are found:

$$w = \frac{\gamma_1}{2}\left(1 - \frac{2\mu}{\gamma_1}\right)^2 \quad ; \quad u_1 = 1 - \frac{2\mu}{\gamma_1} \quad ; \quad m_1 = \sqrt{1 - \left(\frac{2\mu}{\gamma_1}\right)^2} \, . \tag{44}$$

The variance in fitness per site, finally, is proportional to the mean fitness in the PP phase: $V_r = 8\mu w$. Note that all these expressions are independent of the transition rate μ_2 and directly comparable to the results of the two-state model [2, 27] by idebtifying $\{++, +-\}$ with '$+$' and $\{-+, --\}$ with '$-$'.

4.3 Quadratic fitness model: Finite sequence length

For the Fujiyama model with independent sites, all the quantities calculated here, means and variances per site in infinite populations, are independent of the assumed length N of the sequences. This is no longer the case for models including epistasis. In this subsection, we therefore present a quick numerical investigation of the symmetric fitness model for finite system sizes and compare the results with those in the thermodynamic limit. Since the frequencies of genotypes with negative values of the surplus no longer vanish for finite sequences, we use the truncated fitness function (10), with $\gamma_i \equiv \gamma > 0$ and $\alpha_i = 0$ for our calculations.

All results are obtained by direct numerical solution of the eigenvalue problem in the $[(N+1)(N+2)(N+3)/6]$-dimensional vector space of permutation invariant population vectors. Numerically precise calculations have been performed up to $N = 60$ (39711-dim.), the results are shown in Fig. 6. It is seen that the mean surplus and the mean and the variance of the fitness rapidly approach the limiting curves and behave qualitatively different from the Fujiyama model even for very small system sizes. We also show the finite-size behaviour of the variance of the surplus V_s. Since this quantity vanishes as $1/N$, it is not obtainable from the leading order terms in the thermodynamic limit. In our finite size calculations, we rescale V_s with the sequence length to obtain comparable results. Whereas V_s is monotonously increasing for the additive model (where $NV_s = 1 - u^2$), it runs through a maximum for quadratic fitness. Note that this maximum, in contrast to the variance of fitness, is located directly at the error threshold. The behaviour is qualitatively similar to the two-state model [21].

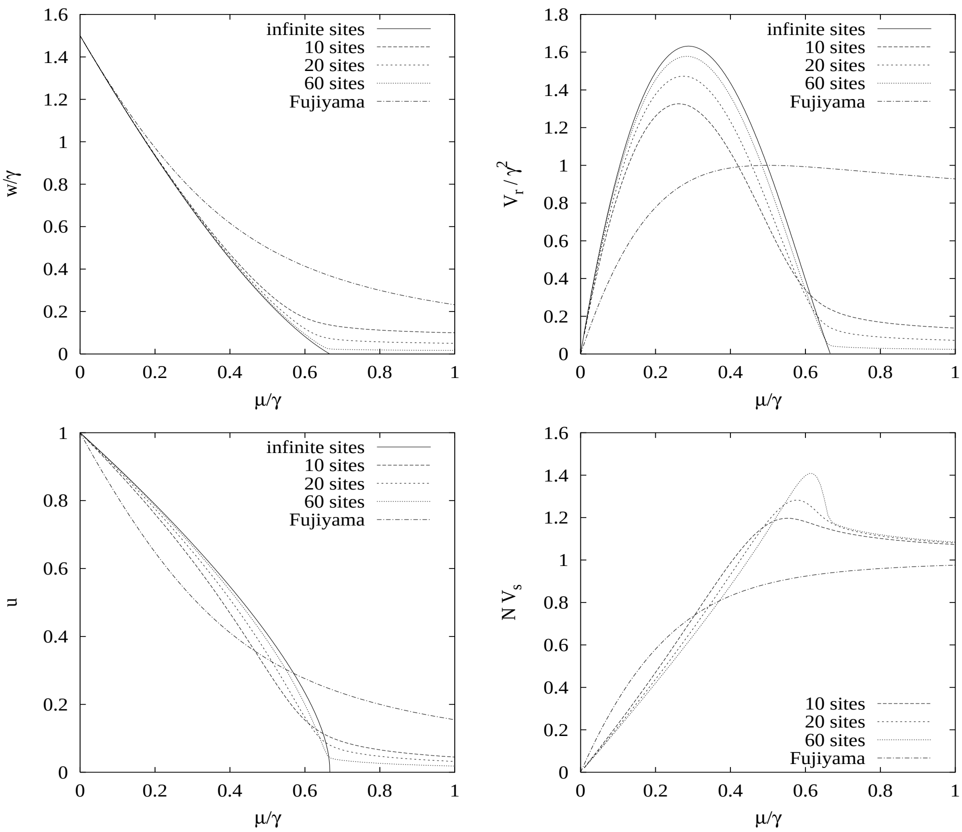

Figure 6: Equilibrium behaviour of fitness and surplus of the symmetric fitness model with finite sequence length. Results for the Fujiyama model with scaling $\alpha = \gamma/2$ are also shown.

Since there has been some discussion recently on the correct scaling of fitness values and mutation rates with the length of the sequence (cf [10, 4]), let us finally remark that the finite size results in this and the next section show that our choice, keeping fitness and mutation rate *per site* fixed, is adequate for all quantities considered here.

4.4 Quadratic fitness model: Time evolution

Originally, the error threshold has been defined as an equilibrium phenomenon (cf [8, 4]): For special forms of the fitness landscape, there is a finite critical value μ_c of the mutation rate beyond which genetic order is no longer maintained by selection. For the four-state model with quadratic fitness, this situation has been discussed above. However, for a suitable fitness function, the threshold is not necessarily connected with high mutation rates. In this subsection, we consider the relaxation of a non-equilibrium population to mutation-selection balance. It turns out that, depending on the starting configuration, an even stronger threshold effect may be observed in the time evolution of the fitness and the surplus for all mutation rates below the critical equilibrium value.

Zero-mutation limit of the transition-transversion model The essence of the threshold phenomenon in the time evolution is already contained in the selection dynamics alone. In a first step, we therefore disregard mutation altogether by working in the zero-mutation limit. Obviously, we then deal with a classical mean-field model on the physical side. As our starting configuration, we choose the completely unstructured population with an equidistribution of genotypes $|\boldsymbol{p}_0\rangle = 4^{-N}|\Omega\rangle$. In this particular situation, some progress is possible also analytically. Noting that

$$\langle\hat{C}\rangle(t) = \frac{\langle\Omega|\hat{C}\exp(t\mathcal{R})|\Omega\rangle}{\langle\Omega|\exp(t\mathcal{R})|\Omega\rangle} = \frac{\mathrm{tr}(\hat{C}\exp(t\mathcal{R}))}{\mathrm{tr}(\exp(t\mathcal{R}))} \tag{45}$$

for any element $\hat{C}$ of the algebra generated by $\{\sigma_i^{(z,0)}, \sigma_i^{(0,z)}\}$, the biological and physical pictures coincide in this case. Using the fitness function of the transition-transversion model with $\gamma_2 = \gamma_3 \equiv \gamma > 0$, we obtain the following implicit equations for the time evolution of the surplus components:

$$u = \frac{\sinh(2\gamma t u)}{\cosh(2\gamma t u) + \exp[-2\gamma_1 t(2u\coth(2\gamma t u) - 1)]} \tag{46}$$

$$u_1 = \frac{\cosh[\gamma t Q(u_1)] - \exp(-2\gamma_1 t u_1)}{\cosh[\gamma t Q(u_1)] + \exp(-2\gamma_1 t u_1)} \tag{47}$$

where

$$Q(u_1) = \sqrt{(1+u_1)^2 - \exp(4\gamma_1 t u_1)(1-u_1)^2}\,. \tag{48}$$

The resulting dynamical phase diagram is shown in Fig. 7. As in the equilibrium situation, there are three phases. Depending on the ratio $\tilde{\gamma}_1 = \gamma_1/\gamma$, the system directly crosses to an ordered phase after a sharply defined waiting time t_c, or performs two consecutive transitions, entering the PP phase in the first one.

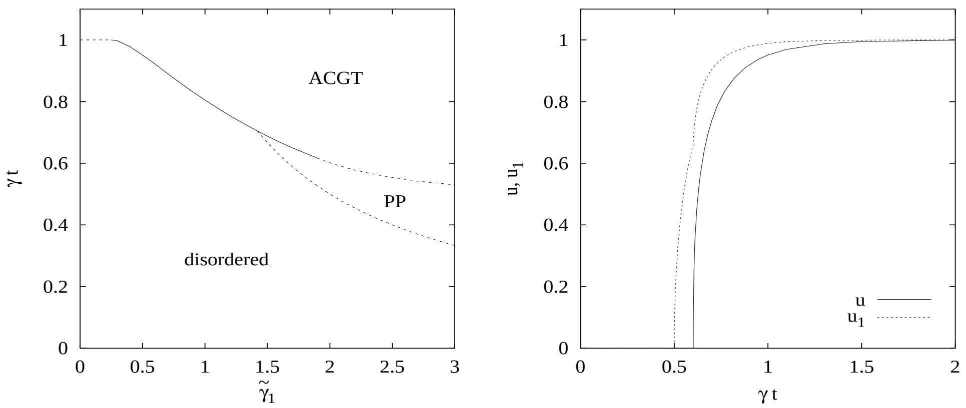

Figure 7: Dynamical phase diagram of the transition-transversion model for vanishing mutation starting from the equidistribution. (Solid: first order; dashed: second order transition). Right: Time evolution of the surplus components for $\tilde{\gamma}_1 = 2$.

As in the equilibrium phase diagram, the dynamical transitions may be of first or second order.

- Second order transitions are located at $\tilde{t} = \gamma t = 1$ for $\tilde{\gamma} \leq 1/4$ and at $\tilde{t} = 1/\tilde{\gamma}_1$ for the transition from the disordered phase to the PP phase. The transition from the PP phase to the ACGT phase is second order above $\tilde{\gamma}_1 \approx 1.9009$ and implicitly given through $2\tilde{t}_c = 1 + \exp[2\tilde{\gamma}_1(\tilde{t}_c - 1)]$. A similar second order transition (with a one-component order parameter) has also been observed in the two-state model [26, 27].
- In an interval around the symmetry point $\gamma_1 = \gamma$, the system possesses a first order transition (in the sense that there is a finite jump in the magnetization). Note that, in contrast to the equilibrium case, also the surplus and even the mean fitness are discontinous on this line, giving rise to a rather pronounced threshold effect in the evolution dynamics (cf. the solid line in Fig. 8 for $\tilde{\gamma} = 1$).

As for the equilibrium values, we also consider the effect of finite sequence lengths on the time evolution. Again, calculations are performed by direct diagonalization of the symmetric fitness model ($\tilde{\gamma} = 1$). Fig. 8 shows how the jump discontinuity in the mean fitness (internal energy) and the delta function singularity in the variance of the fitness (specific heat) are approached by the finite systems. A threshold phenomenon is absent in the time evolution of the Fujiyama model which is also shown in Fig. 8.

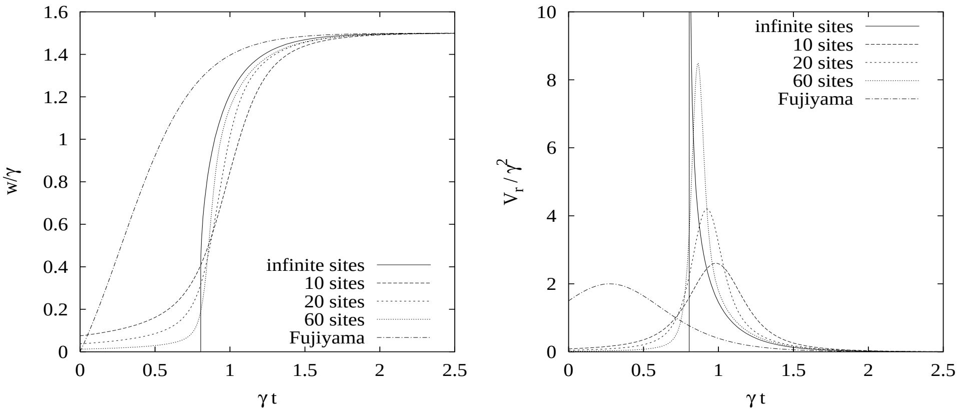

Figure 8: Time evolution of the equidistribution of genotypes in the zero mutation-limit of the symmetric fitness model for different sequence lengths.

Finite mutation rates and different starting configurations In a last step, we now discuss the influence of the mutation rate and the starting configuration on the evolution dynamics. Consider first the time evolution of the equilibrium distribution of genotypes $4^{-N}|\Omega\rangle$. Although no analytical results are available here, we may give the following intuitive argument that there is a phase transition at finite $t = t_c$ for any mutation rate below the critical equilibrium mutation rate μ_c: Since mutation alone tries to keep the population in the equilibrium distribution, the evolution dynamics will be slowed down by mutation for small t. In particular, mean fitness and surplus will remain zero on a finite interval at least up to the threshold value of the corresponding model with vanishing mutation. On the other hand, the limiting values of w and u are finite for $\mu < \mu_c$, giving rise to a non-analytical point of $w(t)$ and $u(t)$ at some finite $t = t_c$. As shown in the upper graph of Fig. 9, this behaviour is clearly visible in numerical results for finite sequence sizes.

In order to contrast the time evolution of the unstructured population with an equidistribution of genotypes as starting configuration, we have also performed

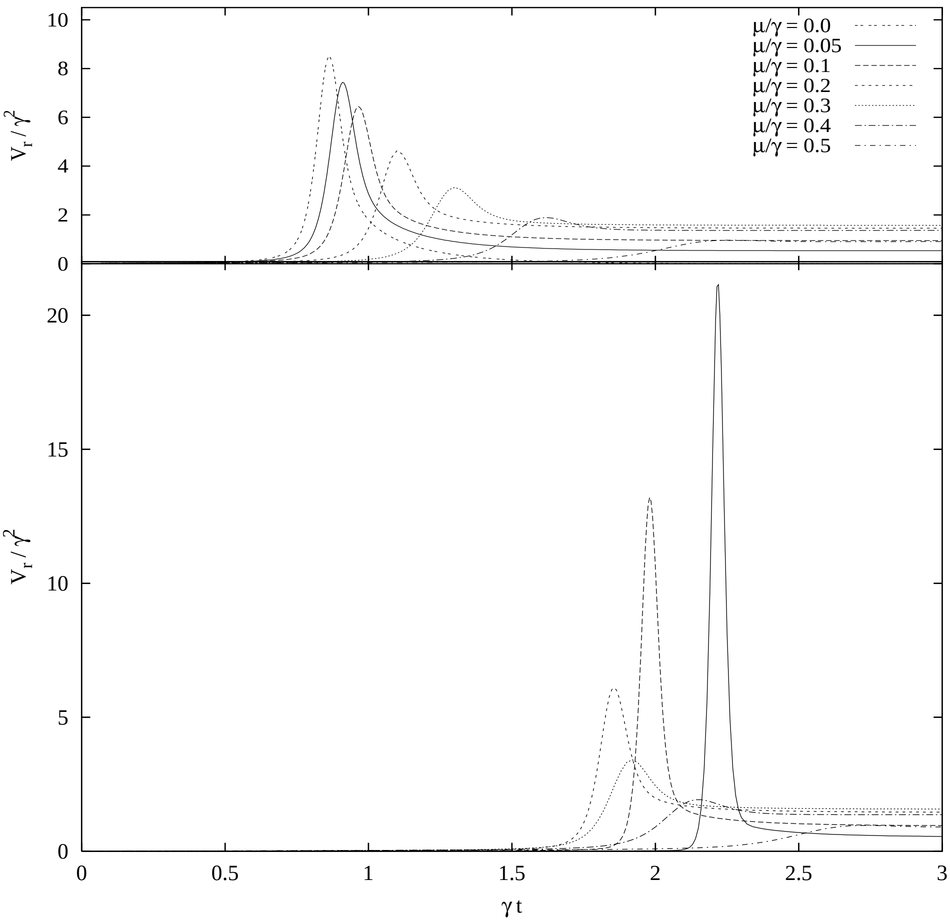

Figure 9: Time evolution of the variance of the fitness in the symmetric fitness model with sequence length $N = 60$. Results are shown for varying mutation rates and two different starting configurations.

calculations for the opposite case of a population with initially homogeneous phenotypes. Here, at $t = 0$, any "individual" in the population has the same value $s_i = 0$ for the three surplus components. The result (for finite sequence length $N = 60$) is shown in the lower viewgraph of Fig. 9. As for the equidistribution of genotypes, there is a clear threshold effect in the time evolution for any finite value $0 < \mu < \mu_c$ of the mutation rate. The transition appears to be particularly sharp for small mutation rates. In contrast to the unstructured case, the critical waiting time t_c for the transition is no longer monotonously increasing with the mutation rate μ, but is separated in two regimes: For mutation rates near the equilibrium threshold value μ_c, the situation is similar to the unstructured case: Here, single mutants with higher fitness appear in the population after a short while. Due to the continuing mutation pressure, however, a certain time is needed for these fitter individuals to grow to a finite proportion and to dominate the mean values in the infinite population. For small μ, on the other hand, the critical waiting time t_c is dominated by the time needed for mutation to explore the configuration space and to generate individuals with higher fitness at a sufficient rate.

5 Discussion

When in [2] a class of models for sequence space evolution was introduced, using the framework of Ising quantum chains, the calculations started with four major simplifications of the biological situation. These are the consideration of a two-state model, the assumption of an infinite sequence length, the use of simplistic fitness landscapes, and the restriction on infinite population sizes. In this paper, we have looked at the first two of these simplifying assumptions. Finally, an extended discussion of the evolution dynamics of these models has also been presented. In the following paragraphs, we give a summary of our findings and an outlook on the remaining open problems.

Two-state versus four-state models. The main concern of this contribution is the generalization of the modelling framework, introduced in [2], to four states (corresponding to the four nucleotides) on each site. The generalization presented makes use of the $C_2 \times C_2$ symmetry inherent in the *Kimura 3 ST* mutation scheme. On the 'physical side' this leads to a model of two coupled Ising quantum chains (rather than to a four-state Potts model). Compared with the two-state model, the extension can be thought of as consisting of two steps. In a first step, we represent the four states on each site by the spin values of two spins in decoupled chains. Note that already in this simplified model three phases occur in the phase diagram since the transition lines of the two decoupled chains will not in general coincide. The second step consists of the introduction of a more realistic mutation scheme which also changes the configuration space topology and the corresponding use of a refined fitness landscape. Both these extensions lead to a coupling of the chains, and an even richer phase space structure is found, including first-order transitions. As may be seen from the introduction of a small linear field term into the fitness function in subsection 4.2, this change of the transition to first order leads to an increased robustness of the threshold phenomena with respect to symmetry-breaking perturbations.

Finite sequence length. Typical sequence lengths of enzymes or viruses are of the order 10^3 – 10^4. While these numbers are certainly far off the typical sizes of macroscopic systems in physics, they are, in principle, large enough to successfully supress $1/N$-corrections. However, especially models with simple fitness landscapes describe – at best – the evolution dynamics in a very restricted configuration space of particularly 'important' sites, disregarding neutral or altogether lethal mutations. In view of this fact, consideration of finite sequence lengths is indispensible and calculations in the thermodynamic limit even seem to be questionable at first sight. In order to clarify the usefulness of infinite-size methods in this context, we performed a number of numerical calculations for finite sequence lengths. The results are quite encouraging. As shown in subsection 4.3, the characteristic properties of the thermodynamic limit are well visible even for tiny sequencs sizes, such as $N = 10$, and the approximation is already quantitatively reasonable for sequences of length 60.

The fitness landscape. The construction of a tractable fitness landscape which nevertheless comprises the relevant biology is certainly the major task for all these models. In this contribution, in order to obtain at least some analytical results, we have chosen a fitness function from the smooth end of the landscape zoo. Due to its permutation invariance, the quadratic fitness function effectively disregards any local variance in the interaction between sites, but only considers the average epistatic effect. As such, it is in many respects certainly no more than a toy-model for evolution. However, the essential structure responsible for characteristic effects such as the error threshold is already contained in this simplified framework. The results obtained here may also serve as a reference for future work on fitness functions with increased ruggedness, such as the NK-landscape hierarchy [13]. Here, we expect the results for the quadratic fitness model to be qualitatively stable at least under certain forms of mild ruggedness, such as the introduction of site-randomness in the fields and interactions [6]. Pronounced changes, on the other hand, should be expected when spin-glass effects come into play.

Finite population size. In going from the deterministic limit to the evolution of finite populations, the ordinary differential equation (2) has to be replaced by the master equation of a stochastic process which is no longer covered by the theoretical framework presented in this article. Due to the complexity of the stochastic equations, analytical results seem to be out of reach at present for all but the simplest selection schemes. Monte-Carlo simulations, however, should be possible and could considerably add to theoretical insight here.

Although the general picture of the deterministic case should persist at least for sufficiently large populations, the study of finite population effects is certainly of importance. For related models, such as the quasispecies model with the *single peaked* landscape, it is has been found [19] that the deterministic results can be interpreted as the time averages of the stochastic process for mutation rates outside a certain interval around an error transition. Directly at the threshold, however, large fluctuations and a jump in the long-time averages appear in the stochastic system at a critical mutation rate which seems to be lower by an amount roughly proportional to $1/\sqrt{N}$ in comparison with the deterministic case. Mainly because of these expected finite population effects we have restricted discussions in this article entirely to the phase space structure of the models and the order of the phase transitions. Any further details of the transitions, even critical exponents, will presumably never be visible in real biological systems and thus seem to be of limited relevance in this context.

Let us finally remark that, although biological populations are certainly finite, the consideration of the infinite population limit is not (only) a technical necessity, but also of direct importance for the study of the error threshold. That is so because this effect, in distinction to the phenomenon of Muller's ratchet, is *by definition* not due to genetic drift, but solely due to the form of the fitness function. It has thus always to be shown that the threshold effect persists even for infinitly large population sizes.

Error threshold behaviour Since there are more than one and sometimes conflicting definitions of the error threshold in literature (cf. the discussion in [4]), let us start this paragraph with a few clarifying remarks. In this article, following [4], we use the notion of the error threshold as equivalent to phase transitions. As such, a clear-cut mathematical definition (as non-analytical points in the mean fitness) is possible only in the infinite sites (or thermodynamic) limit. However, since the thermodynamic limit can be considered as an excellent approximation already for rather small systems, the infinite system property gives a valid explanation for prominent features which are observable for finite sequences as well. In our study, we have always considered sequences of a fixed length and have treated the mutation rate per site as the variable driving the transition. In comparing systems of different length, we have scaled the variables such that a well-defined limit is approached as $N \to \infty$. In particular, the 'critical' mutation rate per site in a finite system quickly converges to the limiting value $\tilde{\mu}_c$. Originally, the threshold has been viewed as a limitating factor on the sequence length [7]. This, however, should not be confusing: We switch to this latter picture simply by letting the reduced mutation rate depend linearly on the sequence length, $\tilde{\mu} \sim N$, and obtain a critical length $N_c \sim \tilde{\mu}_c$ (for sufficiently large sequences).

Our results on the error threshold phenomenon fit previous ones for the two-state case and related models in that negative epistasis is needed to observe a transition (cf. [28, 4]). Contrary to the two-state case, the threshold corresponds to a first-order transition for certain parameter ranges and persists for a sufficiently small linear part in the fitness function. Both, the equilibrium and the dynamical phase diagram of the transition-transversion model (with $\alpha_i = 0$), possess two ordered phases characterized by non-zero values of one or all three components of the surplus order-parameter and the disordered phase with zero surplus where selection ceases to operate. The threshold effect appears to be especially sharp in the evolution dynamics, where a jump in the mean surplus and fitness and a delta singularity in the variance of fitness occurs.

Besides the threshold effect, however, other properties of mutation-selection models may be studied within the framework presented. After all, exclusive concentration on phase transitions is perhaps too much a physicist's point of view on these systems. The relations between surplus, mutation rate and the variance of fitness (24), (36), for example, are valid for the entire time evolution and arbitrary mutation rates. Depending on the fitness function applied, they may give rise to characteristic features also far off the transition point. This is particularly explicit for the equilibrium variance of fitness which runs through a pronounced maximum for fitness functions with negative epistasis at a mutation rate much smaller than the threshold value.

Acknowledgments

It is our pleasure to thank Ellen Baake and Oliver Redner for numerous discussions and comments on the manuscript. Financial support from the German Science Foundataion (DFG) is gratefully acknowledged.

References

[1] E. Baake, Diploid models on sequence space, *J. Biol. Syst.* **3** (1995) 343–9.

[2] E. Baake, M. Baake and H. Wagner, Ising quantum chain is equivalent to a model of biological evolution, *Phys. Rev. Lett.* **78** (1997) 559–62; Erratum: *Phys. Rev. Lett.* **79** (1997) 1782.

[3] E. Baake, M. Baake and H. Wagner, Quantum mechanics versus classical propability in biological evolution, *Phys. Rev.* **E57** (1998) 1191–2.

[4] E. Baake and W. Gabriel, Biological evolution through mutation, selection, and drift: An introductory review, *Ann. Rev. Comput. Phys.* **7** (*in press*, cond-mat/9907372).

[5] J. Crow and M. Kimura, *An Introduction to Population Genetics Theory*, Harper & Row (New York 1970).

[6] N.G. Duffield and R. Kühn, The thermodynamics of site-random mean-field quantum spin systems, *J. Phys.* **A22** (1989) 4643–58.

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

[8] M. Eigen, J. McCaskill and P. Schuster, The molecular quasi-species, *J. Chem. Phys.* **75** (1989) 149–263.

[9] R.A. Fisher, *The Genetical Theory of Natural Selection*, Clarendon Press (Oxford 1930).

[10] S. Franz and L. Peliti, Error threshold in simple landscapes, *J. Phys.* **A26** (1993) 4481–7.

[11] S. Franz, L. Peliti, and M. Sellitto, An evolutionary version of the random energy model, *J. Phys.* **A26** (1993) L1195–9.

[12] S. Galluccio, Exact solution of the quasispecies model in a sharply-peaked landscape, *Phys. Rev.* **E56** (1997) 4526–39.

[13] S.A. Kauffmann and S.A. Levin, Towards a general theory of adaptive walks on rugged landscapes, *J. Theor. Biol.* **128** (1987) 11–45.

[14] J. Kogut, An introduction to lattice gauge theory and spin systems, *Rev. Mod. Phys.* **51** (1979) 656–713.

[15] I. Leuthäusser, An exact correspondence between Eigen's evolution model and a two-dimensional Ising system, *J. Chem. Phys.* **84** (1986) 1884–5.

[16] I. Leuthäusser, Statistical mechanics of Eigen's evolution model, *J. Stat. Phys.* **48** (1987) 343–60.

[17] W.-H. Li, *Molecular Evolution*, Sinauer (Sunderland, 1997).

[18] K. Malarz and D. Tiggemann, Dynamics in Eigen's evolution model, *Int. J. Mod. Phys.* **C9** (1997) 481–90.

[19] M. Nowak and P. Schuster, Error thresholds of replication in finite populations. Mutation frequencies and the onset of Muller's ratchet. *J. Theor. Biol.* **137** (1989) 375–95.

[20] P. O'Brien, A genetic model with mutation and selection, *Math. Biosci.* **73** (1985) 239–51.

[21] O. Redner, *private communication* (1999).

[22] P. Stadler, Landscapes and their correlation functions, *J. Math. Chem.* **20** (1996) 1–45.

[23] D. Swofford, G. Olsen, P. Waddell and D. Hillis, Phylogenetic inference, in: M. Hillis, C. Moritz and E. Mable (Eds.): *Molecular Systematics*, Sinauer (Sunderland, 1995), pp. 407–517.

[24] P. Tarazona, Error thresholds for molecular quasispecies as phase transitions: From simple landscapes to spin-glass models, *Phys. Rev.* **A45** (1992) 6038–50.

[25] C.J. Thompson and J.L. McBridge, On Eigen's theory of the self-organization of matter and the evolution of biological macromolecules, *Math. Biosci.* **21** (1974) 127–42.

[26] H. Wagner, *Biologische Sequenzraummodelle und Statistische Mechanik*, PhD thesis, University of Tübingen, Dissertations Druck (Darmstadt 1998).

[27] H. Wagner, E. Baake and T. Gerisch, Ising Quantum chain and sequence evolution, *J. Stat. Phys.* **92** (1998) 1017–52.

[28] T. Wiehe, Model dependency of error thresholds: the role of the fitness functions and contrasts between the finite and infinite sites models, *Genet. Res. Camb.* **69** (1997) 127–36.