RANDOM MATRICES AND LOCALIZATION IN THE QUASISPECIES THEORY*

B. Waclaw

School of Physics and Astronomy, University of Edinburgh
Mayfield Road, Edinburgh EH9 3JZ, United Kingdom

(Received April 11, 2011)

The quasispecies model of biological evolution for asexual organisms such as bacteria and viruses has attracted considerable attention of biological physicists. Many variants of the model have been proposed and subsequently solved using the methods of statistical physics. In this paper I will put forward important but yet overlooked relations between localization theory, random matrices, and the quasispecies model. These relations will help me to study the dynamics of this model. In particular, I will show that the distribution of times between evolutionary jumps in the genotype space follows a power law, in agreement with recent findings in the shell model — a simplified version of the quasispecies model.

DOI:10.5506/APhysPolB.42.1141
PACS numbers: 87.23.Kg, 87.10.–e, 02.50.–r, 05.70.Fh

1. Introduction

The theory of biological evolution is a pillar of modern biology and has been inspiring scientists since the time of Darwin. Evolution operates by two primary processes: natural selection, in which best adapted organisms outcompete less adapted ones, and genetic drift which is the source of variation in relative frequencies of different traits within a population of individual organisms. Although natural selection was supported by rather strong experimental evidence already 150 years ago, it was only in 1950s when the discovery of the role of DNA in heredity and, subsequently, the explanation of its molecular structure provided a molecular basis for genetic variability. Today we know that this variability is related to changes to the genetic material stored in DNA, either by DNA exchange in sexual reproduction or microbial conjugation, or by mutations — random changes in the sequence of nucleotides which form DNA strands.

* Presented at the XXIII Marian Smoluchowski Symposium on Statistical Physics, "Random Matrices, Statistical Physics and Information Theory", Kraków, Poland, September 26–30, 2010.

All organisms capable of self-reproduction which exist today are complicated systems of many coupled chemical reactions, usually taking place in isolated compartments — cells — or even smaller regions within cells. Is evolution a phenomenon which started to operate after living organisms had emerged a few billion years ago, or was it preceded by an analogous process acting on molecules floating freely in oceans? In an attempt to answer this question, Eigen and Schuster conceived a theoretical model [1] explaining how selection and genetic drift could work already on the level of single chemical molecules. In their model, macromolecules such as DNA or RNA, which are linear polymers composed of nucleotides, were subjected to error-prone replication. If errors (mutations) were infrequent, the molecule with the highest replication rate soon dominated the population. For increasing mutation rate, however, the fittest molecule became surrounded by an increasing cloud of mutants in the space of all possible sequences. Eigen and Schuster called this cloud "quasispecies", by analogy to the concept of biological species, which consists of closely related genotypes.

The model predicts that if the mutation rate increases beyond some critical value — an error threshold — the quasispecies becomes "delocalized". This means that the fittest molecule corresponding to some sequence of nucleotides is lost and all possible sequences start to appear. If these other sequences are much less fit or even incapable of reproduction (lethal mutations), the population will inevitably die out. This is called an error catastrophe. The error threshold is predicted to decrease with increasing length of sequences, suggesting that for a given mutation rate, the amount of genetic information stored in a self-replicating molecule is restricted. On the other hand, higher mutation rate means improved adaptability to changing conditions. Indeed, it has been found experimentally [2], that the evolution of some viruses such as HIV operates very close to the error threshold.

The molecular quasispecies theory has attracted considerable attention not only from biologists but also mathematicians and physicists. In particular, the quasispecies model has been studied by mapping it onto Ising spin chains [3, 4, 5, 6], directed polymers [7], or, more recently, Anderson localization [8, 9], and has been solved exactly in some special cases [5, 10, 11]. The purpose of this work is to re-examine the quasispecies model, or, more precisely, its version with parallel mutation and selection [5], and to show an analogy between this model and random matrices which appear in Anderson localization. I will first define the quasispecies model and discuss its several versions which appear in the literature. Next, I will show how some questions fundamental to the quasispecies theory can be rephrased using the language of random matrix theory (RMT). Finally, I will show how to answer these questions using some simple concepts borrowed from RMT.

2. Quasispecies model

The quasispecies model [1] describes biological evolution of simple organisms which reproduce asexually in a chemostat — a bioreactor with constant supply of fresh nutrients and removal of liquid culture and waste products, so that the culture volume is kept constant. Every organism has a "genotype" which is a sequence of length L composed of symbols taken from some finite alphabet. The symbols could correspond to four different nucleotides which are the building blocks of DNA, but one usually considers binary sequences composed of only two symbols 0 and 1. This assumption simplifies calculations but it has no qualitative influence on any properties of the model I will mention in this paper. The number of all possible genotypes is $N = 2^L$. Every genotype can be labelled by an integer number $i = 1, \ldots, N$, and the binary representation of $i - 1$ gives the corresponding binary sequence.

The genotypes replicate by copying themselves. However, the copy procedure is not error-free: each symbol has a finite probability γ of being substituted by another (randomly chosen) symbol in the process of replication. Assuming that errors are made independently at any position in the sequence, the matrix Q_{ji} which describes the probability of mutation from genotype i to genotype j reads

$$Q_{ji} = \gamma^{d(i,j)}(1-\gamma)^{L-d(i,j)}\,, \tag{1}$$

where $d(i,j)$ is the Hamming distance between the genotypes i, j, *i.e.*, $d(i,j)$ is the number of positions at which the corresponding sequences are different. For $j = i$, $Q_{ii} = (1-\gamma)^L$ gives the probability of replication without errors. By definition, Q_{ji} is a symmetric, square, doubly stochastic matrix: $\sum_j Q_{ji} = \sum_i Q_{ij} = 1$.

If we denote by ϕ_i the specific growth rate of genotype i, we can describe the model by the following reactions

$$\text{genotype}\, i \xrightarrow{\phi_i} \begin{cases} \text{genotype}\, i + \text{genotype}\, i\,, & \text{(with probability } Q_{ii}) \\ \text{genotype}\, i + \text{genotype}\, j\,, & \text{(with probability } Q_{ji}) \end{cases} \tag{2}$$

in which ϕ_i above the arrow means that the reaction constant is ϕ_i. Suppose now that the population is very large so that we can neglect stochastic fluctuations of the numbers of genotypes. Then, the time evolution of the abundances (number densities) $n_1, \ldots, n_N$ of genotypes $i = 1, \ldots, N$ is modelled by the set of N differential equations:

$$\frac{d}{dt} n_i(t) = \sum_{j=1}^{N} Q_{ij}\phi_j n_j(t) - n_i(t)J(t)\,, \tag{3}$$

where $J(t)$ is the rate at which organisms (molecules) are washed out from the system. The term $J(t)$ forces the system to evolve towards the steady state, otherwise the growth would always be exponential. In this paper, $J(t)$ is assumed to be proportional to the overall concentration $\sum_i n_i(t)$. This causes the net growth rates to become negative if the population is too dense, as if the organisms competed for limited resources.

Since $J(t)$ depends on $n_i(t)$, the quasispecies equation (3) is non-linear. However, a simple change of variables

$$x_i(t) = n_i(t) \exp\left(\int_0^t J(t')dt' \right) \tag{4}$$

reduces Eq. (3) to a linear equation

$$\frac{d}{dt} x_i(t) = \sum_{j=1}^{N} Q_{ij}\phi_j x_j(t) \, . \tag{5}$$

As we are interested only in the ratios of different n_is (relative concentration of genotypes), we can use x_i instead of n_i to describe the state of the system. The above equation can be rewritten as

$$\frac{d}{dt} \vec{x}(t) = W \vec{x}(t) \, , \tag{6}$$

where the matrix $W_{ij} = Q_{ij}\phi_j$. Let $\lambda_1 > \lambda_2 > \cdots > \lambda_N$ be the set of eigenvalues of W, and $\{\vec{\psi}_i\}$ denote the corresponding eigenvectors:

$$W\vec{\psi}_i = \lambda_i \vec{\psi}_i \, . \tag{7}$$

Then, Eq. (6) has the following solution

$$\vec{x}(t) = e^{tW}\vec{x}(0) = \sum_i e^{t\lambda_i} \left(\vec{\psi}_i \cdot \vec{x}(0) \right) \vec{\psi}_i \, , \tag{8}$$

which for large times reduces to $\vec{x}(t \to \infty) \propto \vec{\psi}_1$. Therefore, the steady state of Eq. (3) is proportional to the eigenvector $\vec{\psi}_1$ to the largest eigenvalue λ_1 of the matrix W. This gives the steady-state abundances

$$\vec{n}^* \equiv \vec{n}(t \to \infty) = \frac{\lambda_1}{\sum_i \psi_{1,i}} \vec{\psi}_1 \, . \tag{9}$$

The physical properties of the formal solution from Eq. (8) and Eq. (9) depend on the choice of the growth rates $\{\phi_i\}$, which specify "fitnesses" of different genotypes, *i.e.*, how well they are adapted to the environment. The graph of possible mutations plus the fitnesses is usually referred to as the fitness landscape. This metaphor is based on viewing fitness peaks as mountains of different heights, with the population climbing generally uphill in the course of evolution and moving from lower to higher peaks, until the highest peak (global fitness maximum) is reached. Although frequently used in population biology, the concept of fitness has had an important drawback: until recently little information was available on real fitness landscapes because experimental evaluation of the fitness for large numbers of genotypes is very difficult. Lacking experimental data, many models for the fitness landscape have been considered, without *a priori* knowledge of which one is correct. Some of the most popular choices are listed below:

1. Single-peak landscape [10]: the fitness ϕ_1 of a single genotype is taken to be maximal, and $\phi_2 = \cdots = \phi_N$ are assumed to be smaller than ϕ_1. This corresponds to a situation in which there is one best-fit genotype (master sequence) and all mutants are less fit.

2. Multiplicative landscape [11]: ϕ_1 is maximal, and $\phi_i = \phi_1(1-s)^{d(1,i)}$ where s is some positive constant and $d(1,i)$ is the Hamming distance between the sequences 1 and i. In this model, each single-symbol mutation lowers the fitness by the same amount.

3. "Holey" landscape: the fitness of each genotype is either large (fit genotypes) or small (unfit genotypes). Thus, in the fitness landscape, there are holes of unfit genotypes surrounding islands of equally fit genotypes. The fitnesses can be either correlated [12] or uncorrelated [13].

4. Rugged fitness landscape [14, 15]: the fitnesses are drawn as independent, identically distributed random numbers from some continuous distribution $p(\phi)$, the same for all genotypes.

Most analytical results were obtained for the single-peaked landscape (1), whereas the most realistic one is probably the rugged landscape (4). However, what all these landscapes have in common is the emergence of localized "quasispecies" and the existence of error threshold.

As already mentioned, the quasispecies is a set of closely related sequences which occupy a finite area in the genotype space. The name "quasispecies" comes from an apparent similarity to a real-world situation in which genotypes corresponding to organisms of the same species form a "cloud" of mutants in the genotype space around the best-fit genotype. The quasispecies appears for any mutation rate γ smaller than some critical γ_c, because in this case the steady state solution $\vec{n}^* \propto \vec{\psi}_1$ is localized around (usually) the maximal fitness, see Fig. 1. Above γ_c, the steady state becomes delocalized and spreads over the whole genotype space. The critical

$\gamma_{\rm c}$ can be easily calculated [2] in the case of a single-peak fitness landscape and reads

$$\gamma_{\rm c} = 1 - \left(\frac{\phi_2}{\phi_1}\right)^{1/L}, \tag{10}$$

where ϕ_1, ϕ_2 are fitnesses of the best-fit genotype and less-fit genotypes, respectively. Above $\gamma_{\rm c}$, the best-fit genotype (the master sequence) vanishes from the population. This critical $\gamma_{\rm c}$ is called the error threshold and the transition from localized to delocalized quasispecies is known as the error catastrophe. For increasing L and ϕ_2/ϕ_1 kept constant, the critical mutation rate scales as $\gamma_{\rm c} \sim 1/L$, thus longer sequences have smaller error thresholds.

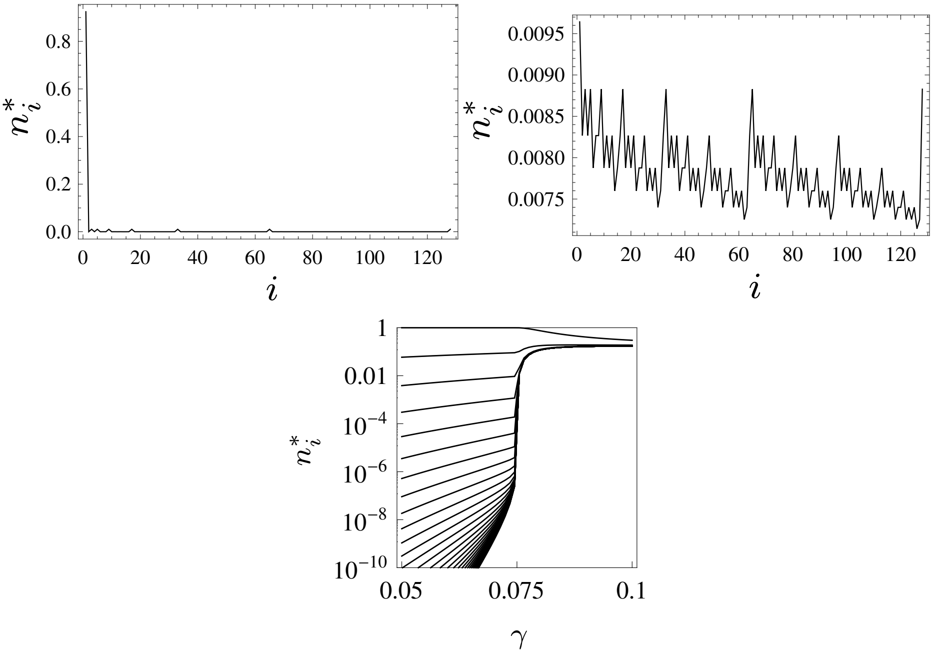

Fig. 1. Abundances n_i^* of genotypes in the steady state in the single-peak model, for $L = 7$ (which corresponds to $N = 128$), $\phi_1 = 10, \phi_2 = \cdots = \phi_N = 1$, thus the fittest sequence being at $i = 1$, and for two mutation rates $\gamma = 0.01$ (top left, well below the error threshold $\gamma_{\rm c}(L = 7) \approx 0.28$) and $\gamma = 0.4$ (top right, above the threshold). Genotypes which correspond to the same Hamming distance from the master sequence $i = 1$ (the same "error class") have the same abundance. Bottom: plot of n_i for different error classes (lines from top to bottom) as a function of γ, for $L = 30$. The transition, which is clearly visible at $\gamma_{\rm c}(L = 30) \approx 0.074$, becomes sharper for increasing L.

3. Quasispecies model with parallel mutation and selection

The quasispecies model has an even simpler counterpart — para-mu-se (parallel mutation and selection) model [5]. A key feature of this model is that, in comparison to equations (2), growth and mutation are decoupled:

$$\text{genotype } i \xrightarrow{\phi_i-\gamma\sum_j A_{ij}} 2 \text{ genotype } i\,, \tag{11}$$

$$\text{genotype } i \xrightarrow{\gamma A_{ij}} \text{genotype } i + \text{genotype } j\,. \tag{12}$$

Here A is the adjacency matrix of a graph of possible mutations and γ is the mutation rate. It is usually assumed that the matrix A is symmetric (forward and reverse mutation have the same probability) and that $A_{ij} = 1$ if the Hamming distance $d(i,j) = 1$. Therefore, mutations are allowed to change at most one symbol per replication. The matrix A is the adjacency matrix of a hypercube graph, see Fig. 2.

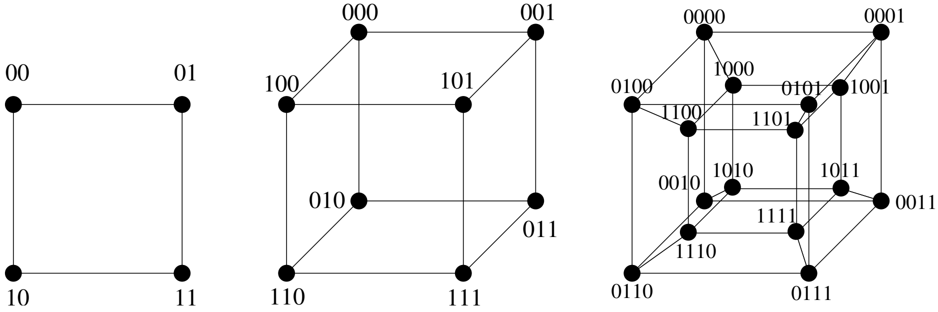

Fig. 2. Examples of mutations graphs for the para-mu-se model, for $L = 2$, 3, 4.

The model is known to have the error threshold for some fitness landscapes [6][1] and for small γ (for which $Q_{ij} \approx \gamma A_{ij}$ in Eq. (1)) it behaves very much the same as the quasispecies model from Section 2. However, the para-mu-se version of the quasispecies model is generally simpler from a mathematical point of view. From now on, I will focus on this model, although general conclusions of this work remain valid also for the original quasispecies model.

In the limit of infinite populations, the para-mu-se model is described by the following set of equations (*cf.* Eq. (3)):

$$\frac{d}{dt}n_i(t) = n_i(t)(\phi_i - J(t)) + \gamma \sum_{j=1}^{N} A_{ij}(n_j(t) - n_i(t))\,. \tag{13}$$

[1] For the single-peak landscape, eigenvector ψ_1 is always localized, but the net growth rate of the master sequence can be negative above some critical γ_c, which may be thought as a sort of error catastrophe.

Applying the same transformation (4) as before, the above set is reduced to the linear form

$$\frac{d}{dt}\vec{x}(t) = W\vec{x}(t)\,, \tag{14}$$

where the matrix W is now defined as $W_{ij} = \delta_{ij}\phi_i + \gamma\mathit{\Delta}_{ij}$, and $\mathit{\Delta}_{ij} = A_{ij} - \delta_{ij}\sum_k A_{ik}$ is the graph Laplacian. If $\lambda_1 > \lambda_2 > \cdots > \lambda_N$ and $\{\vec{\psi}_i\}$ are again eigenvalues and eigenvectors of W, the solution of Eq. (14) is given by the same Eq. (8) as for the quasispecies model, but with the new matrix W.

4. Localization, random matrices and quasispecies

The quasispecies or the para-mu-se model can be solved exactly in some special cases. The purpose of this article is to show that in the case of random, rugged fitness landscape, which is much more difficult to study, there is an interesting analogy between the para-mu-se model, Anderson localization and RMT. Surprisingly, this analogy (however obvious it may seem) is not widely appreciated.

In what follows I will thus assume that $\{\phi_i\}$ are random numbers taken from some distribution $p(\phi)$. The quasispecies matrix W,

$$W = \begin{bmatrix} \phi_1 & 0 & \dots & 0 \\ 0 & \phi_2 & \dots & 0 \\ & & \ddots & \\ 0 & \dots & 0 & \phi_N \end{bmatrix} + \gamma\mathit{\Delta} = D + \gamma\mathit{\Delta}\,, \tag{15}$$

is a sum of a diagonal, random matrix $D = \mathrm{diag}(\phi_1, \ldots, \phi_N)$, and a Laplacian matrix $\gamma\mathit{\Delta}$. The eigenproblem of W:

$$\sum_j (\phi_i\delta_{ij} + \gamma\mathit{\Delta}_{ij})\psi_j = \lambda\psi_i \tag{16}$$

can be translated to the following Schroedinger equation:

$$-\sum_j \mathit{\Delta}_{ij}\psi_j + V_i\psi_i = E\psi_i\,, \tag{17}$$

where $V_i = -\phi_i/\gamma$ and $E = -\lambda/\gamma$. This is precisely the Anderson model of localization on arbitrary lattices (see, *e.g.*, Ref. [16]) with random potential V_i. High fitness values correspond to low potential values, and the ground state corresponds to the steady state $\vec{n}^* \propto \vec{\psi}_1$ (the quasispecies).

The matrix W has random elements only on the diagonal, which makes it different from what is usually considered to be a random matrix — a matrix with a finite fraction of elements being (possibly correlated) random numbers. Such "dense" random matrices, which form the core of random matrix theory, appear in many problems in physics [17], telecommunication and information theory [18], and quantitative finance [19]. However, matrices in which only diagonal elements (or elements in a narrow band) are random, are also quite abundant in physics, in particular in quantum chaos [20,21,22] and Anderson localization problems [23, 24, 25, 26, 27]. The matrix W with random fitness values defined above belongs to the same class of sparse random matrices. There is, however, one important difference: systems studied in the framework of localization theory are usually low-dimensional, except for Bethe lattices [16,23]. The reason is that low-dimensional systems can model real physical situations like transport properties of disordered solids [28]. In addition, many analytical results have been obtained for 1d or Bethe lattices due to their special, simple structure. In contrast, the quasispecies problem is multi-dimensional, since the Laplacian $\varDelta$ is defined on the hypercube graph, and the matrix W (albeit sparse) has a non-trivial structure. Although this can generally make analytical calculations very hard, it will not be relevant for the problems studied in this work.

Random matrix theory deals primarily with eigenvalues and eigenvectors of matrices. Table I lists some of typical quantities calculated in RMT. It turns out that some of them are directly related to quantities relevant to the quasispecies theory. The first example I shall consider are spectral properties of the matrix W. A simple reasoning shows that differences between nearest eigenvalues of W determine typical time scales in the model. In the limit of small mutation rate $\gamma \approx 0$, W is almost diagonal and the eigenvalues $\{\lambda_i\}$ are equal to the fitnesses $\{\phi_i\}$. For simplicity, let the fitnesses decrease with i, so that the ordered eigenvalues are $\lambda_i \cong \phi_i$. All eigenvectors are then trivially localized: $\psi_{i,j} \cong \delta_{ij}$, and the best adapted genotype corresponds to the eigenvector $\vec{\psi}_1$, the second best adapted one to the eigenvector $\vec{\psi}_2$, and so on. If the population is initially localized at the least adapted genotype N, then scalar products $\vec{\psi}_i \cdot \vec{x}(0)$ decay exponentially fast with increasing Hamming distance $d(N,i)$. Writing the solution of Eq. (14) as

$$\begin{aligned}\vec{x}(t) &= e^{tW}\vec{x}(0) = \sum_i e^{t\lambda_i}\left(\vec{\psi}_i\vec{x}(0)\right)\vec{\psi}_i = e^{t\lambda_1}\left[\left(\vec{\psi}_1\vec{x}(0)\right)\vec{\psi}_1 + e^{-t(\lambda_1-\lambda_2)}\right. \\ &\quad \times \left.\left[\left(\vec{\psi}_2\vec{x}(0)\right)\vec{\psi}_2 + e^{-t(\lambda_2-\lambda_3)}\left[\left(\vec{\psi}_3\vec{x}(0)\right)\vec{\psi}_3 + \dots\right]\right]\right]\end{aligned}$$

reveals that the contribution of eigenvectors $\vec{\psi}_N, \vec{\psi}_{N-1}, \dots, \vec{\psi}_2$ first increase (in this order) and then decay with rates $\lambda_{N-1} - \lambda_N, \dots, \lambda_2 - \lambda_3, \lambda_1 - \lambda_2$. The rates $\lambda_i - \lambda_{i+1}$ give different time scales in the system. In particular,

$$\tau = \frac{1}{\lambda_1 - \lambda_2} \tag{18}$$

is the characteristic time to reach the steady state $\vec{n}^* \propto \vec{\psi}_1$. Other differences correspond to characteristic times

$$\tau_i = \frac{1}{\lambda_i - \lambda_{i+1}} \tag{19}$$

related to "jumps" between locally adapted quasispecies (Fig. 3, left). If the distribution of differences $\lambda_i - \lambda_{i+1}$ is known, one can calculate the average time between these events, and hence estimate the speed of evolution. However, the probability distribution of differences $S_n(s)$ with $s = \lambda_n - \lambda_{n+1}$ is just the nearest-neighbour spacing distribution, which is very commonly used in RMT to study short-range fluctuations in the spectrum (Fig. 3, right). Therefore, the problem of time evolution in the quasispecies theory is equivalent to the problem of finding $S_n(s)$ for a particular ensemble of random matrices W. The rest of the paper is devoted to calculating this quantity and interpreting it from a quasispecies perspective.

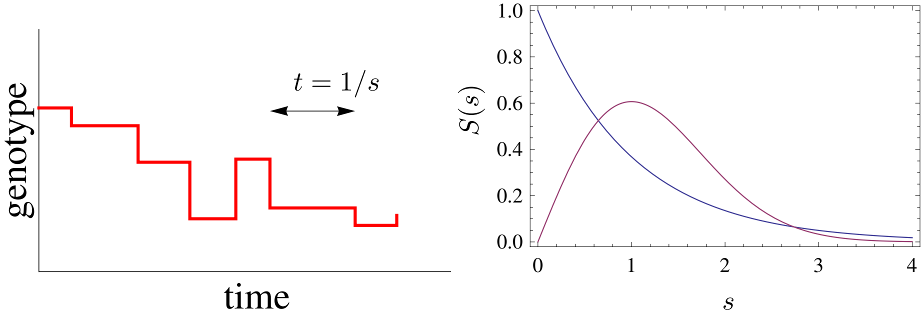

Fig. 3. Left: a schematic view of jumps in the fitness space. The red line shows the trajectory of the most populated genotype. Typical time scale t is inversely proportional to the separation s of eigenvalues of W. Right: two level spacing distributions — exponential and Wigner's surmise — frequently encountered in RMT.

Before I proceed to calculations of $S_n(s)$, I will briefly mention other similarities between the quasispecies theory and RMT. For example, the formula for the steady-state abundances $\vec{n}^*$ from Eq. (9) shows that the maximal eigenvalue λ_1 plays the role of the total abundance of all possible genotypes. The distribution of the maximal eigenvalue is frequently studied in the framework of RMT.

Finally, the statistics of eigenvectors $\{\vec{\psi}_i\}$ of some random matrices turns out to be very important for localization problems. In particular, participation ratio of eigenvectors is studied as a measure of localization. However, the localization transition in matrix models corresponds to the error catastrophe in the quasispecies model, which shows yet another interesting connection between RMT and the quasispecies theory. Table I provides a summary of all analogies mentioned in this work.

TABLE I

Correspondence between quantities of interest in RMT and in the quasispecies theory.

RMT		Quasispecies
level spacing	↔	statistics of jumps in fitness space
distribution of maximal eigenvalue	↔	steady-state total abundance
localization of eigenvectors	↔	error threshold
participation ratio of eigenvectors	↔	genetic diversity

5. Level spacing distribution

In this section I will discuss the level spacing distribution $S_n(s)$, and its average over all nearest-neighbour pairs of eigenvalues

$$S(s) \equiv \frac{1}{N-1} \sum_{n=1}^{N-1} S_n(s) \tag{20}$$

for the matrix W in the para-mu-se model. I will assume the uniform distribution of fitness: $p(\phi) = 1$ in the range $0 \ldots 1$. This ensures that there is always a maximal and a minimal fitness, which is biologically relevant (the growth rate cannot be arbitrarily large), and that for large $N = 2^L$, $\max\{\phi_1, \ldots, \phi_N\} \to 1$ is bounded from above[2]. At last, the uniform distribution of ϕ_i allows one to draw yet another link to Anderson localization, in which site potentials are also uniformly distributed [30].

As explained above, dynamical properties of the quasispecies or para-mu-se model can be inferred from the distribution $S_n(s)$. One of the most important results of RMT is that eigenvalues usually "repel" each other in the spectrum [17] so that $S(s)$ is zero at $s = 0$. In particular, for Gaussian random matrices, we have to a good approximation

$$S(s) \approx s e^{-s^2/2}\,, \tag{21}$$

which is the well-known Wigner surmise. This level repulsion is characteristic for interacting systems, in which eigenvalues are correlated. For uncorrelated eigenvalues, the level-spacing distribution is exponential, $S(s) = e^{-s}$,

[2] Note that for $p(\phi) = e^{-\phi}$ which is often assumed by various authors [29, 15], $\max\{\phi_i\} \sim L$.

for unfolded spectrum, *i.e.*, after transforming all eigenvalues such that their spectral density is uniform. Both the Wigner surmise and the exponential distribution are plotted in Fig. 3, right.

It is evident from Eq. (15) that the mutation rate γ plays the role of interaction strength, so we expect that $S(s)$ should be exponential for $\gamma \to 0$, and that it will show signs of level repulsion for $\gamma > 0$. This is indeed seen in Fig. 4, in which I plotted $S(s)$ obtained from numerical diagonalization of W for uniform fitness distribution and various mutation rates. A simple argument shows that for L large enough,

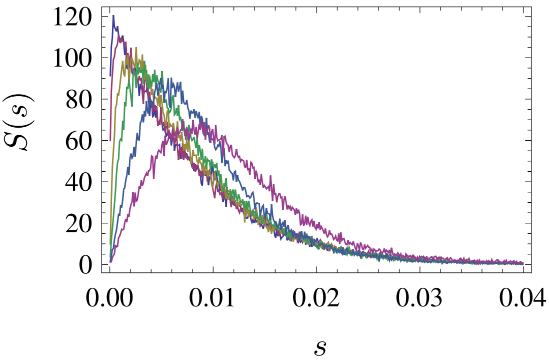

Fig. 4. Level-spacing distribution $S(s)$ for $L = 7$ and $\gamma = 0.005$, 0.01, 0.02, 0.03, 0.05, 0.1 (curves from left to right.)

$$S(s) \cong N f(\gamma L, Ns)\,, \tag{22}$$

where $f(g, x)$ is a (yet unknown) semi-positive function. This means that $S(s)$ obtained for different lengths L and for mutation rates $\gamma = g/L$ with some arbitrary g should collapse to a single curve when plotted in the rescaled variable $x = Ns$, *i.e.*, when we "blow up" the spectrum of eigenvalues. This can indeed be seen in Fig. 5. The scaling form (22) has a simple motivation. Firstly, the number of eigenvalues of the matrix W is N. Since $W = D + \gamma \Delta$, for small γ we expect that the spectrum of W will have a similar width as the spectrum of D. But the spectrum $\rho_D(\lambda)$ of the diagonal matrix D reads $\rho_D(\lambda) = p(\lambda)$, where $p(\phi) = 1$ for $0 \leq \phi \leq 1$, and it has a finite support of length one. Therefore, the average distance between the eigenvalues of W must scale as $1/N$, which explains the factor N blowing up the nearest-level spacing in Eq. (22).

Secondly, the net growth rate of genotype i is $\phi_i - L\gamma + \gamma \sum_j A_{ij} n_j / n_i = \phi_i + O(\gamma L)$, thus γ appears in the para-mu-se equations (13) always as a product of L and γ. We thus suspect that γL is the relevant variable for the balance between growth and mutation. To make it more explicit, we

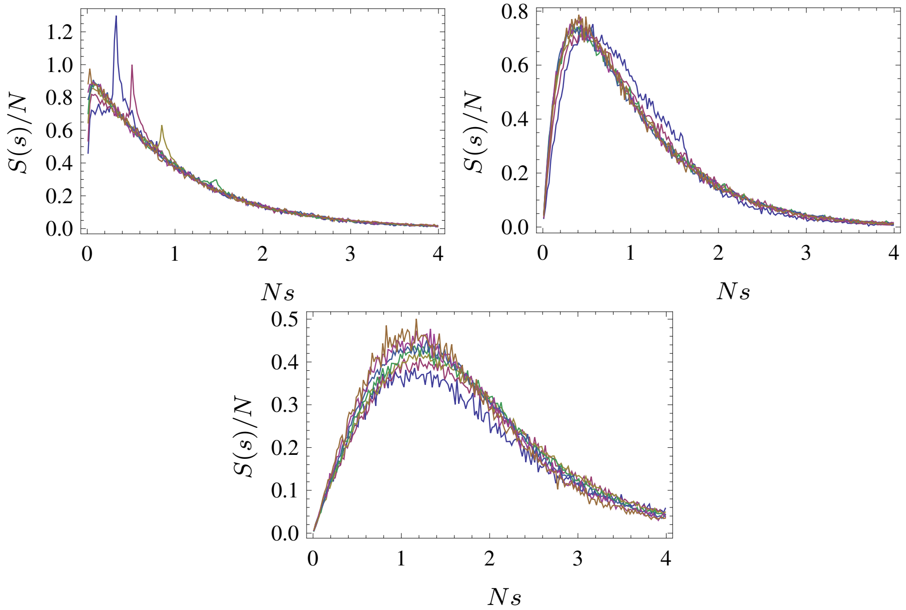

Fig. 5. Rescaled $S(s)$ for $L = 4$, 5, 6, 7, 8, 9, 10, and $\gamma = g/L$, where $g = 0.04$ (top left), 0.2 (top right) and 1 (bottom). For $g = 0.04$ some spikes are visible for $s = 2\gamma N$. These spikes are finite-size effects, which vanish with increasing L.

observe that for small γ, the quasispecies is localized at the largest fitness, which for the uniform distribution $p(\phi)$ is $\phi_1 \approx 1$. This best adapted genotype is surrounded by a sea of less adapted genotypes with average fitness $\phi_2 \approx 1/2$. The relative abundances of the best adapted genotype, x_1, and a less-adapted one, x_2, can be approximately determined from

$$\dot{x}_1(t) = x_1(t)(\phi_1 - L\gamma) + L\gamma x_2(t)\,, \tag{23}$$

$$\dot{x}_2(t) = x_2(t)\phi_2 + \gamma x_1(t)\,. \tag{24}$$

Here I have assumed that mutations of the less-adapted genotype produce mainly genotypes from the same, less-adapted class of genotypes, therefore there is no loss term due to mutations in the second equation. Then, the difference of eigenvalues of the corresponding 2×2 matrix W reads

$$\lambda_1 - \lambda_2 = \sqrt{L^2\gamma^2 + (\phi_1 - \phi_2)^2 + 2L\gamma(2\gamma - \phi_1 + \phi_2)}\,, \tag{25}$$

and it is evident that (assuming that $2\gamma \ll \phi_1 - \phi_2 \approx 1/2$) $s = \lambda_1 - \lambda_2$ depends only on the product of $L\gamma$ in the limit of small mutation rate.

The N-scaling from (22) can be deduced analytically for uniform $p(\phi)$, and for $\gamma \to 0$. In this limit, as already mentioned, the eigenvalues of W are distributed uniformly between 0 and 1. This means that, effectively, there are no interactions in the system, so $S(s)$ for the unfolded spectrum should be exponential. This is very easy to check. The probability $S_n(s)$ that the difference for an ordered pair of eigenvalues $\lambda_n > \lambda_{n+1}$ will be s is given by

$$S_n(s) = Z^{-1} \int\limits_{-\infty}^{\infty} d\lambda_1 \int\limits_{\lambda_1}^{\infty} d\lambda_2 \cdots \int\limits_{\lambda_{N-1}}^{\infty} d\lambda_N p(\lambda_1) \cdots p(\lambda_N) \delta(\lambda_{n+1} - \lambda_n - s)\,, \tag{26}$$

where

$$Z = \int\limits_{-\infty}^{\infty} d\lambda_1 \int\limits_{\lambda_1}^{\infty} d\lambda_2 \cdots \int\limits_{\lambda_{N-1}}^{\infty} d\lambda_N p(\lambda_1) \cdots p(\lambda_N)\,. \tag{27}$$

Equations (26) and (27) can be evaluated for the uniform fitness distribution. We obtain $Z = 1/N!$ and

$$S_n(s) = N(1-s)^N \approx Ne^{-Ns}\,, \tag{28}$$

which does not depend on n, thus the average level-spacing is also $S(s) \cong Ne^{-Ns}$. For large N, $S(s)$ scales as in Eq. (22) with $\gamma = 0$ and

$$f(x) = e^{-x}\,. \tag{29}$$

Such an exponential decay is indeed visible in Fig. 5, top left.

Numerical simulations presented in Fig. 5 indicate that the scaling form (22) is valid also for $\gamma \equiv g/L > 0$, *i.e.*, when γ scales inversely with the length L of the sequence. For g small enough, the same scaling holds for the distribution $S_1(s)$ of the gap between two largest eigenvalues $s = \lambda_1 - \lambda_2$, see Fig. 6, top left. However, for large g, the distributions shift to larger s with increasing L (Fig. 6, top right and bottom). Figure 7 shows plots of the participation ratio divided by the total number of genotypes,

$$\mathrm{PR} = \frac{1}{2^L} \left(\sum_i \psi_{1,i} \right)^2 \Big/ \sum_i \psi_{1,i}^2\,, \tag{30}$$

which is the measure of localization: PR ≈ 0 for eigenvectors with only few entries larger than zero, whereas PR ≈ 1 means that all entries are roughly the same. In Fig. 7, PR ≈ 0 for small g, but for sufficiently large g, PR is of order one. This means that the principal eigenvector covers a finite fraction of the genotype space, hence the quasispecies is no longer localized. This indicates a transition similar to the error catastrophe in the quasispecies model.

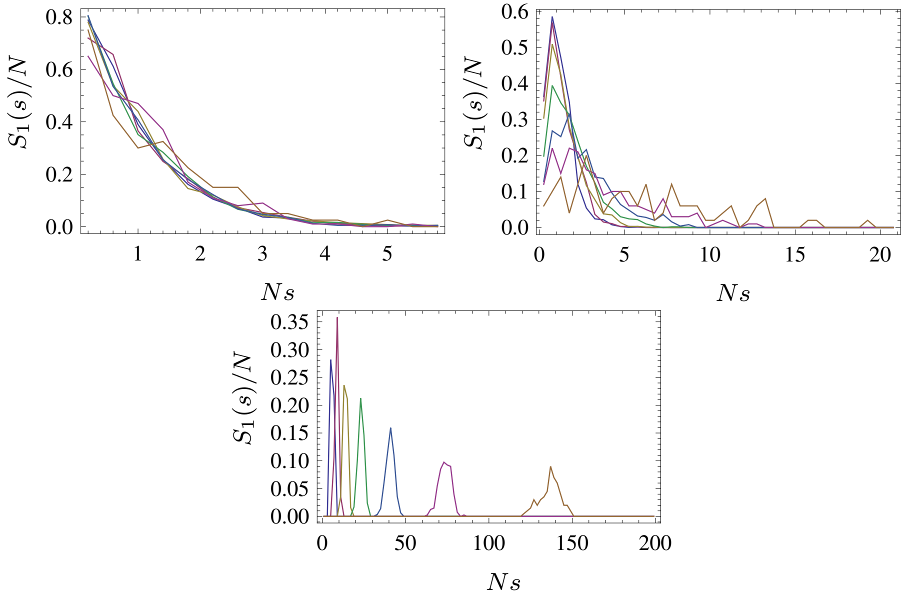

Fig. 6. Rescaled distribution of the gap between two largest eigenvalues $S_1(s)$, for $L = 4, \ldots, 10$, $\gamma = g/L$, and $g = 0.04$ (top left), $g = 0.2$ (top right) and $g = 1.0$ (bottom). The distribution shifts to the right for increasing L.

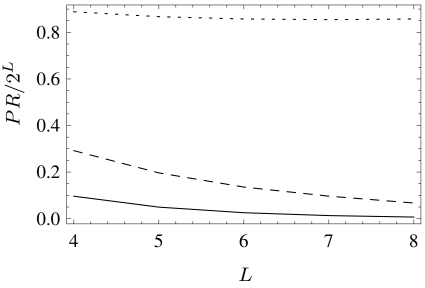

Fig. 7. Participation ratio divided by $N = 2^L$ for $g = 0.04$ (solid), $g = 0.2$ (dashed) and $g = 1.0$ (dotted).

6. Statistics of jumps

Using the scaling hypothesis for the level-spacing distribution which, as we have seen, is valid for any $\gamma > 0$ small enough (so that the quasispecies remains localized), the statistics of jumps can be inferred very easily. The distribution $P_{\text{jumps}}(\tau)$ of times between jumps is given by

$$P_{\text{jumps}}(\tau) = \int_0^\infty S(s)\delta(\tau - 1/s)ds = N\frac{1}{\tau^2}f(\gamma L, N/\tau)\,. \qquad (31)$$

The same formula holds for $P_{\rm ss}(\tau)$, the distribution of times to steady state, because the gap distribution $S_1(s) \approx S(s)$ for localized quasispecies. This means that the mean time to steady state, which is the time it takes until the quasispecies stops evolving, is

$$\langle\tau\rangle = N \int\limits_0^\infty \frac{1}{x} f(\gamma L, x)\, dx\,, \tag{32}$$

and it grows linearly with N, or exponentially with the length of the sequence L. For $\gamma = 0$, the above integral is divergent, because $f(\gamma L, 0) > 0$. This is correct, because in the absence of mutations there is no way to reach the steady state from any other point in the genotype space. But for $\gamma > 0$, we have $f(\gamma L, 0) = 0$ due to level repulsion. For small x, $f(\gamma L, x)$ is proportional to x as it can be seen in Fig. 5. Therefore, $\int_0^\infty \frac{1}{x} f(\gamma L, x) dx < \infty$, and the average time $\langle\tau\rangle$ is finite.

Equation (31) tells us that for $N \ll \tau \ll N/\gamma$, for which $f(\gamma L, x) \approx$ const., the distribution of times between jumps follows a power law: $P_{\rm jumps}(\tau) \sim \tau^{-2}$. This is also true for the time to reach steady state, $P_{\rm ss}(\tau) \sim \tau^{-2}$. In Fig. 8 I show plots of cumulative distributions of times between jumps measured directly by solving differential equations (13) numerically for 5000 random fitness landscapes, and tracing the position of the maximal abundance n_i. A jump was recorded whenever this maximal abundance changed its location in the genotype space. In this way, the statistics of times between jumps was obtained. For each fitness landscape, the simulation was stopped when the difference $\sum_i |n_i^* - n_i(t)|/N$ between the steady-state solution and $n_i(t)$ was smaller than 10^{-6}. In this way, also the statistics of

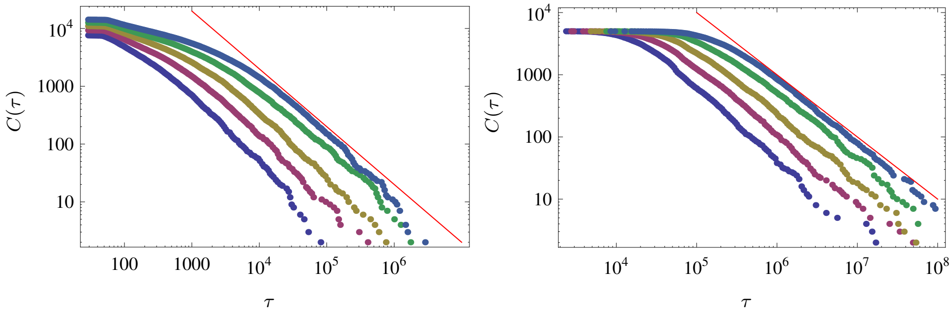

Fig. 8. Plots of the cumulative distribution of jumps $C_{\rm jumps}(\tau) = \int_\tau^\infty P_{\rm jumps}(\tau')d\tau'$ for $L = 4, 5, 6, 7, 8$ (from left to right) and $\gamma = 0.04/L$ (left panel). Solid line corresponds to theoretical result $C(\tau) \sim \tau^{-1}$. $C_{\rm ss}(\tau)$ for the time to reach steady state (right panel).

times to steady state was collected. The experimental cumulative distributions $C_{\rm jumps}(\tau) = \int_\tau^\infty P_{\rm jumps}(\tau')d\tau'$ and $C_{\rm ss}(\tau) = \int_\tau^\infty P_{\rm ss}(\tau')d\tau'$ presented in Fig. 8 show a power-law behaviour with an exponent close to minus one, $C_{\rm jumps}(\tau) \sim C_{\rm ss}(\tau) \sim \tau^{-1}$, in good agreement with theory.

The power-law behaviour of $P_{\rm jumps}(\tau)$ for individual jumps as well $P_{\rm ss}(\tau)$ for the time to reach steady state has been observed in a simplified "shell" model in the strong selection limit [29,31], which would correspond to $\gamma \to 0$ in our model. We see that a simple observation, which relates the dynamics of the quasispecies to the level spacing distribution, allows one to extend this result to the case of $\gamma = g/L > 0$.

7. Conclusion

The main objective of this work was to present an interesting analogy between the quasispecies theory (in particular, the para-mu-se model) and random matrices which appear in localization theory. I discussed how static and dynamical properties of the quasispecies model are related to quantities such as nearest-level spacing distribution or participation ratio of eigenvectors, which are typically calculated within the framework of random matrix theory. Although most of the results presented here were obtained in numerical simulations, they were all corroborated by simple, mathematical calculations. It remains a challenge to calculate analytically the level-spacing distribution for non-zero mutation rates.

The analogy to Anderson localization mentioned here has been already mentioned in Ref. [8] and, more recently, in Ref. [9], in which some results of localization theory for 1d tight-binding models are used to find the point of the phase transition in a model with two quasispecies linked by migration. However, no systematic studies of localization on the hypercube with random distribution of fitness (site potential) have been made so far. This would be potentially a very interesting research area which could further link biological evolution models, localization theory, and random matrices.

I thank Z. Burda, M.R. Evans, and J.-M. Luck for many valuable discussions. I am particularly indebted to R. Allen for introducing me to problems of biological (especially microbial) evolution. I also acknowledge support by the EPSRC under grant EP/E030173.

REFERENCES

[1] M. Eigen, P. Schuster, *Naturwiss.* **64**, 541 (1977).

[2] M.A. Nowak, *Trends in Ecology and Evolution* **7**, 118 (1992).

[3] I. Leithäusser, *J. Stat. Phys.* **48**, 343 (1987).

[4] L. Demetrius, *J. Chem. Phys.* **87**, 6393 (1987).

[5] E. Baake, H. Wagner, *Genet. Res.* **78**, 93 (2001).

[6] E. Baake, M. Baake, H. Wagner, *Phys. Rev. Lett.* **78**, 559 (1997).

[7] E. Kussel, S. Leibler, A. Grosberg, *Phys. Rev. Lett.* **97**, 068101 (2006).

[8] C.L. Epstein, *J. Stat. Phys.* **124**, 25 (2006).

[9] B. Waclaw, R.J. Allen, M.R. Evans, *Phys. Rev. Lett.* **105**, 268101 (2010).

[10] S. Galluccio, *Phys. Rev.* **E56**, 4526 (1997).

[11] G. Woodcock, P.G. Higgs, *J. Theor. Biol.* **179**, 61 (1996).

[12] S. Gavrilets, *Am. Nat.* **154**, 1 (1999).

[13] S. Gavrilets, J. Gravner, *J. Theor. Biol.* **184**, 51 (1997).

[14] S.A. Kauffman, S. Levin, *J. Theor. Biol.* **128**, 11 (1987).

[15] K. Jain, J. Krug, *Genetics* **175**, 1275 (2007).

[16] R. Abou-Chacra, D.J. Thouless, P.W. Anderson, *J. Phys.* **C6**, 1734 (1973).

[17] T. Guhr, A. Müller-Groeling, H.A. Weidenmüller, *Phys. Rep.* **299**, 189 (1998).

[18] A.L. Moustakas *et al.*, *Science* **287**, 287 (2000).

[19] L. Laloux, P. Cizeau, J.-P. Bouchaud, M. Potters, *Phys. Rev. Lett.* **83**, 1467 (1999).

[20] G. Casati, L. Molinari, F. Izrailev, *Phys. Rev. Lett.* **64**, 1851 (1990).

[21] A.D. Mirlin, Y.V. Fyodorov, *J. Phys.* **A26**, L551 (1993).

[22] K. Życzkowski, M. Lewenstein, M. Kuś, F. Izrailev, *Phys. Rev.* **A45**, 811 (1992).

[23] G. Biroli, G. Semerjian, M. Tarzia, *Phys. Rev.* **B80**, 014524 (2009).

[24] F.L. Metz, I. Neri, D. Bolle, *Phys. Rev.* **E82**, 031135 (2010).

[25] E. Gudowska-Nowak, G. Papp, J. Brickmann, *J. Phys. Chem.* **A102**, 9554 (1998).

[26] P. Neu, R. Speicher, *J. Stat. Phys.* **80**, 1279 (1995).

[27] G. Biroli, P. Monasson, *J. Phys.* **A32**, L255 (1999).

[28] B. Kramer, A. MacKinnon, *Rep. Prog. Phys.* **56**, 1469 (1993).

[29] J. Krug, Ch. Karl, *Physica A* **318**, 137 (2003).

[30] P.W. Anderson, *Phys. Rev.* **109**, 1492 (1958).

[31] K. Jain, *Phys. Rev.* **E76**, 031922 (2007).