

The tangled nature model as an evolving quasi-species model

Simone Avogadro di Collobiano¹, Kim Christensen¹ and
Henrik Jeldtoft Jensen²

¹ Blackett Laboratory, Imperial College London, Prince Consort Road, London SW7 2BW, UK

² Department of Mathematics, Imperial College London, 180 Queen's Gate, London SW7 2BZ, UK

E-mail: k.christensen@imperial.ac.uk and h.jensen@imperial.ac.uk

Received 20 August 2002, in final form 7 November 2002

Published 15 January 2003

Online at stacks.iop.org/JPhysA/36/883

Abstract

We show that the tangled nature model can be interpreted as a general formulation of the quasi-species model by Eigen *et al* in a frequency-dependent fitness landscape. We present a detailed theoretical derivation of the mutation threshold, consistent with the simulation results, that provides a valuable insight into how the microscopic dynamics of the model determines the observed macroscopic phenomena published previously. The dynamics of the tangled nature model is defined on the microevolutionary time scale via reproduction, with heredity, variation and natural selection. Each organism reproduces at a rate that is linked to the individuals' genetic sequence and depends on the composition of the population in genotype space. Thus the microevolutionary dynamics of the fitness landscape is regulated by, and regulates, the evolution of the species by means of the mutual interactions. At low mutation rate, the macroevolutionary pattern mimics the fossil data: periods of stasis, where the population is concentrated in a network of coexisting species, are interrupted by bursts of activity. As the mutation rate increases, the duration and the frequency of bursts increase. Eventually, when the mutation rate reaches a certain threshold, the population is spread evenly throughout the genotype space showing that natural selection only leads to multiple distinct species if adaptation is allowed time to cause fixation.

PACS numbers: 87.10.+e, 87.23.-n, 87.23.Kg

1. Introduction

Explaining the observed macroevolutionary patterns as collective emergent properties of systems with many interacting degrees of freedom, whether these be single individuals or

'species', is an alluring challenge for researchers with a background in statistical physics [1, 2]. The quasi-species model by Eigen *et al* [3, 4] has proved useful when investigating the behaviour of populations in a given fixed fitness landscape and it provides a firm paradigm for many models [5].

The fundamental idea in the approach by Eigen *et al* is to identify species with sequences in genotype space. The positions in genotype space which are assigned particularly high fitness are called wildtypes, that is, the forms that predominate in a population are well adapted to the environment. During the reproduction event, mutations are seen as errors of the replication of the parental sequence. The effect is thus to spread the population from the original point to neighbouring positions in genotype space. If one were to use a classical Darwinian view on such a process, the population would then be sharply localized in genotype space on the position which corresponds to a high fitness: all other positions would be cancelled by their low fitness. This can only be true if the replication process is by and large accurate. A replication process with a too high mutation rate would produce copies of the original fit parent with so many errors that selection is unable to maintain the population at the original point.

By lowering the mutation rate progressively, variation would be less effective in dispersing the population since the offspring are more similar to the parents. The quasi-species model defines the presence of a threshold, in the mutation rate, where the multiplication process changes drastically. The gradual decrease of the mutation rate sees the transition from a random population, diffused as scattered points in genotype space, to a population constrained to a few positions.

The transition of the process from a random state to an ordered one is a phase transition, with the mutation rate acting as a control parameter. The nature of the transition has been extensively studied. In the seminal paper by Eigen *et al*, where the transition was first noticed [4], the species have a predetermined fixed fitness associated. Subsequently, the quasi-species model has been analysed in different fitness landscapes [6, 7], for different topologies of the genotype space [8], and for spatially resolved models [9], each confirming the original results. Finally, the error threshold in a model with a dynamical fitness landscape [10] has been analysed. In this case however, the dynamics is regulated artificially from outside.

It has been shown that it is possible to map the original quasi-species model onto a two-dimensional Ising system with nearest-neighbour interaction in one direction [11], and that, in this representation, for simple fitness landscapes, the correspondence links the error threshold with a first-order phase transition [12]. A relation of fundamental importance by Galluccio *et al* [13] proves that the error threshold naturally arises as a consequence of the model introduced and that, more generally, for a given mutation rate p_{mut} and a given reproduction rate p_{off} it is possible to determine uniquely an upper limit for the length of the genetic sequence.

We show that the tangled nature model, introduced in detail in [14, 15], can be considered as a general formulation of the quasi-species model. The generalization is provided by a relaxation on the condition of fixed population size, which, in the original formulation, acts as selection principle on the sequences. The most important feature of the tangled nature model, for details see [14, 15], is that of creating multiple co-evolving quasi-species in a frequency-dependent fitness landscape, where the dynamics of the landscape is an inherent property of the model. In this paper, we present in detail the theoretical calculation of the mutation threshold which fits the experiment accurately [15].

It is also interesting to point out the connection between the tangled nature model and game theoretical nonlinear replicator dynamics [16]. In both cases the reproduction of a given type of individuals depends on the configuration of the entire population. One therefore

expects to find stable solutions to the dynamics of the tangled nature model similar to the Nash equilibria or evolutionary stable strategies found for replicator dynamics. We have stressed this relation by using the term quasi-evolutionary stable strategies to denote the quasi-stable configurations of the tangled nature model.

In section 2, we review briefly the quasi-species model by Eigen *et al.* Section 3 briefly discusses the definition of the tangled nature model with an intrinsically generated dynamic fitness landscape. We discuss in detail the dynamics of the model in terms of difference equations. Section 4 contains a discussion of the error threshold theoretically and numerically and finally in section 5 we discuss the relation between the tangled nature model and the quasi-species model by Eigen *et al.*

2. The quasi-species model

Eigen *et al* [4] introduced a model in which the effects of various mutation rates on a process of replication of finite sequences of binary values are explored. Each sequence $\mathbf{S}^a = \{S_1^a, S_2^a, \dots, S_L^a\}$, where $S_i^a \in \{-1, +1\}$, $i = 1, 2, \dots, L$, in genotype space represents a species. Each existing sequence \mathbf{S}^a replicates, with a constant rate $p_{\text{off}}^a = p_{\text{off}}(\mathbf{S}^a)$ and degrades with a constant and universal (i.e., independent of position) rate p_{kill} . The number $n_a(t) = n(\mathbf{S}^a)(t)$ of copies of a given sequence \mathbf{S}^a varies with time. The replication process is not exact but prone to error. During the replication, the rate of mutation per gene is p_{mut} .

The model has been solved analytically in the limit where one particular sequence is assumed to have a high fitness, while mutants are less fit. For low mutation rate, the population is concentrated around the top of the mountain in the fitness landscape. The dominant sequence with its surrounding mutants is called a *quasi-species*. As the mutation rate increases, the population drifts away from the top down to the ridges. Eventually, when the mutation rate reaches a threshold value $p_{\text{mut}}^{\text{th}}$, the population is spread evenly throughout the fitness landscape, that is, a phase transition occurs at $p_{\text{mut}}^{\text{th}}$.

3. The dynamics of the tangled nature model

The dynamics of the tangled nature model is defined via an elementary time step where (a) one organism is randomly selected and killed with constant probability p_{kill} , (b) one organism is randomly selected and with probability p_{off} , that depends on the current composition of the population in genotype space, two offspring are reproduced and the parent is then removed from the ecology [14, 15].

By analysing the dynamics it is possible to characterize the stable configurations that may develop in the tangled nature model.

The difference equation describing the variation of the number of individuals of a position \mathbf{S}^a during a single time step can be derived as follows. Let $n_a(t)$ denote the number of individuals at position \mathbf{S}^a . Then

$$n_a(t+1) = n_a(t) + \sum_E \Delta n_a(E) P(E) \quad (1)$$

where t is the number of time steps and E refers to any event that can affect n_a , that is, an annihilation event or reproduction event, by an amount $\Delta n_a(E)$. The event E occurs with probability $P(E)$.

For a killing event, $\Delta n_a(E) = -1$ and the probability of a killing event is the product of the probability of choosing an organism in position \mathbf{S}^a times the killing rate, that is,

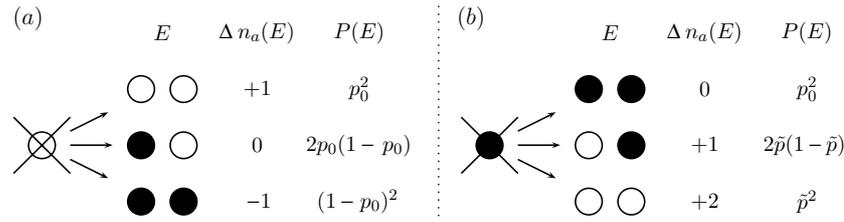


Figure 1. Probabilities associated with a reproduction event. An organism at position \mathbf{S}^a is shown with an open circle and any other type of organism with a solid circle. The columns labelled ‘ E ’ represent the three possible outcomes of a reproduction event; in the columns labelled by ‘ $\Delta n_a(E)$ ’ the variation of n_a associated with event E is listed. The probabilities involved are given in the columns marked $P(E)$, where p_0 is the probability of no mutations during a reproduction event and $1-p_0$ the probability of at least one mutation while \tilde{p} is defined in equation (6). (a) Reproduction originating from \mathbf{S}^a . (b) Evaluation of the backflow associated with the events $\mathbf{S} \neq \mathbf{S}^a \rightarrow \mathbf{S}^a$.

$P(E) = \rho_a(t)p_{\text{kill}}$, where we have introduced the density $\rho_a(t) = \frac{n_a(t)}{\sum_a n_a(t)}$ of organisms at position \mathbf{S}^a .

For a reproduction event, distinction has to be made between the case where reproduction originates from position \mathbf{S}^a , see figure 1(a) and reproduction originating from any other position \mathbf{S}^b different from \mathbf{S}^a , which we will call the ‘backflow’ contribution, see figure 1(b).

The first case happens with probability $P = \rho_a(t)p_{\text{off}}^a(t)$, that is, the probability of picking an organism of position \mathbf{S}^a times the fitness of \mathbf{S}^a . In this event, n_a can decrease by one unit ($\Delta n_a = -1$), increase by one unit ($\Delta n_a = +1$) or remain constant ($\Delta n_a = 0$), with relative probabilities as calculated in figure 1(a).

The probability of having i mutations during a single replication is

$$p_i = \binom{L}{i} p_{\text{mut}}^i (1-p_{\text{mut}})^{L-i} \quad \forall i = 0, 1, \dots, L \quad \text{with} \quad \sum_{i=0}^L p_i = 1. \quad (2)$$

From figure 1(a) we can deduce the net contribution to the population at position \mathbf{S}^a by summing over all possible events:

$$\sum_E \Delta n_a(E) P(E) = p_0^2 - (1-p_0)^2 = 2p_0 - 1. \quad (3)$$

The ‘backflow’ contribution occurs with probability

$$\sum_{b \neq a} \rho_b(t) p_{\text{off}}^b(t). \quad (4)$$

In this case, the variations and the probabilities involved are shown in figure 1(b).

In order to mutate from \mathbf{S}^b to \mathbf{S}^a , Ld_{ab} mutations are necessary, where

$$d_{ab} = d(\mathbf{S}^a, \mathbf{S}^b) = \frac{1}{2L} \sum_{i=1}^L |S_i^a - S_i^b| \quad (5)$$

so

$$\tilde{p} = p_{\text{mut}}^{Ld_{ab}} (1-p_{\text{mut}})^{L(1-d_{ab})} \quad (6)$$

is the probability of creating an organism in position \mathbf{S}^a originating from position \mathbf{S}^b .

As Ld_{ab} mutations are needed, the probability involved in a backflow contribution from position \mathbf{S}^b is, see figure 1(b),

$$\sum_E \Delta n_a(E) P(E) = 2\tilde{p}(1-\tilde{p}) + 2\tilde{p}^2 = 2\tilde{p} = 2p_{\text{mut}}^{Ld_{ab}} (1-p_{\text{mut}})^{L(1-d_{ab})}. \quad (7)$$

Thus, the full expression for the difference equation is

$$n_a(t+1) = n_a(t) - \rho_a(t)p_{\text{kill}} + \rho_a(t)p_{\text{off}}^a(t)(2p_0 - 1) + 2 \sum_{b \neq a} \rho_b(t)p_{\text{off}}^b(t)p_{\text{mut}}^{Ld_{ab}}(1 - p_{\text{mut}})^{L(1-d_{ab})}. \quad (8)$$

This is the equivalent of the quasi-species equation by Eigen *et al.* The main difference is that the rates of production depend on the current composition in population space.

Summing equation (8) over all positions in genotype space we find, as expected,

$$N(t+1) = N(t) - p_{\text{kill}} + \langle p_{\text{off}} \rangle. \quad (9)$$

From the simulations we know that in the limit of strong interactions among the individuals, the dynamics is intermittent [14, 15]. Extended periods are dominated by a network of few heavily occupied positions. These periods, called quasi-evolutionary stable strategies (q-ESS), are interrupted by sharp bursts where the configuration of the species changes rapidly and significantly. In order to describe the dynamics, we impose a stability condition on the difference equation: we require that within a single q-ESS, the average number of individuals remains constant. Moreover, the q-ESS states are dominated by some very fit positions surrounded by unfit neighbouring positions. Thus we can neglect the backflow contribution in the difference equation, equation (8), and obtain

$$n_a(t+1) = n_a(t) + \rho_a(t)[p_{\text{off}}^a(t)(2p_0 - 1) - p_{\text{kill}}]. \quad (10)$$

Averaging over time, the equation becomes

$$\bar{n}_a = \bar{n}_a + \overline{\rho_a p_{\text{off}}^a}(2p_0 - 1) - \overline{\rho_a} p_{\text{kill}}. \quad (11)$$

Assuming that $\overline{\rho_a p_{\text{off}}^a} = \overline{\rho_a} p_{\text{off}}^a$, the fitness for all positions in the set $\mathcal{S}_{p_{\text{off}}=p_q}$:

$$\overline{p_{\text{off}}^a} = \frac{p_{\text{kill}}}{2p_0 - 1} \equiv p_q. \quad (12)$$

With $p_{\text{mut}} = 0.008$ we have $p_0 = (1 - p_{\text{mut}})^L = 0.852$ for $L = 20$; using $p_{\text{kill}} = 0.2$, we find $p_q = 0.284$ consistent with the observation in figure 2.

Neglecting the backflow is valid if all terms

$$\rho_b(t)p_{\text{off}}^b(t)p_{\text{mut}}^{Ld_{ab}}[1 - p_{\text{mut}}]^{L(1-d_{ab})} = \rho_b(t)p_{\text{off}}^b(t)p_{\text{mut}}^{Ld_{ab}}[1 - L(1 - d_{ab})p_{\text{mut}} + \dots]$$

are small. Since $p_{\text{mut}} \ll 1$, the leading term is $\rho_b(t)p_{\text{off}}^b(t)p_{\text{mut}}^{Ld_{ab}}$. This can be neglected if $Ld_{ab} > 1$. With $Ld_{ab} = 1$ it can be neglected since none of the nearest neighbours is fit, as $p_{\text{off}}^b(t) \ll 1$.

4. The error threshold

The discussion of the q-ESS state was made with the implicit assumption of the existence of q-ESS states. We will find here that we can establish qualitative arguments that ensure the existence of the q-ESS states.

We have seen that q-ESS states are possible only if the interactions are important to the weight function. Furthermore, the average fitness p_q of the fit positions in the q-ESS state is given by

$$p_q = \frac{p_{\text{kill}}}{2(1 - p_{\text{mut}})^L - 1} \quad (13)$$

and thus is related to the mutation rate. This relation states that the fit positions are those that are able to counterbalance the killing by the production of offspring.

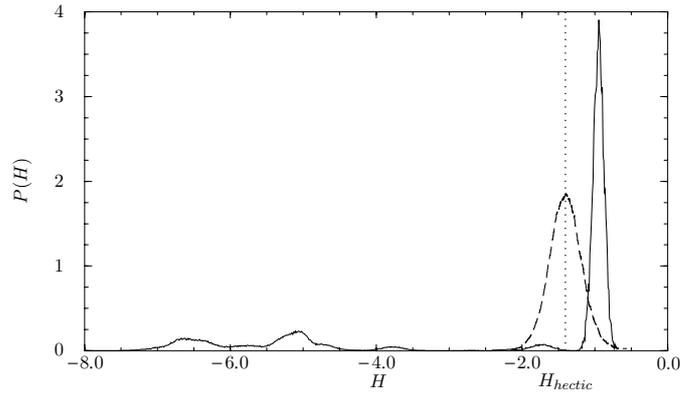


Figure 2. The probability density function of the weight function $H = \ln\left(\frac{p_{\text{off}}}{1-p_{\text{off}}}\right)$ during a q-ESS state of a simulation (solid line) and during a transition between two q-ESS states (dashed line). During a q-ESS state (solid line) positions range in two sets: unfit positions, for which the weight function is lower than -3.0 and fit positions, for which the fitness is greater than the average value $\langle H \rangle = \ln\left(\frac{1-p_{\text{kill}}}{p_{\text{kill}}}\right) \approx -1.38 = H_{\text{hectic}}$, indicated by a vertical dotted line. During a transition (dashed line) the fitness of all positions is normally distributed around H_{hectic} where all positions reproduce (on average) at the same rate, equal to the killing rate. Note the support of the weight function in the hectic phase exceeds H_q , ensuring that the positions in genotype space are able to fulfil the q-ESS balance equation (13). The parameters (for precise definitions, see [14, 15]) are $p_{\text{kill}} = 0.2$, $\mu = 1/1000 \cdot \ln\left(\frac{1-p_{\text{kill}}}{p_{\text{kill}}}\right) \approx 0.0014$, $C = 10.0$ and $p_{\text{mut}} = 0.008$.

Equation (13) is the starting point for determining a necessary condition for the existence of a q-ESS state. We have investigated the behaviour of the dynamics as a function of mutation rate. The results are illustrated in figure 3.

For increasing p_{mut} , the duration of q-ESS states decreases. Above a threshold $p_{\text{mut}}^{\text{th}}$ of the mutation rate p_{mut} , there are no more q-ESS states: the dynamics is completely hectic. For intermediate values of p_{mut} , the transitions between two q-ESS states are extended and the initial transient can be very long.

This numerical result shows that the model defines an error threshold for the mutation rate above which no q-ESS state exists.

From equation (13) we obtain for the weight function

$$H_q = \ln\left(\frac{p_q}{1-p_q}\right) = \ln\left(\frac{p_{\text{kill}}}{2p_0 - 1 - p_{\text{kill}}}\right). \quad (14)$$

When the mutation rate is close to $p_{\text{mut}}^{\text{th}}$, most of the simulations are in hectic states, for which the fitness is equal to p_{kill} and therefore we might assume that the weight function is equal to

$$H_{\text{hectic}} = \ln\left(\frac{p_{\text{kill}}}{1-p_{\text{kill}}}\right). \quad (15)$$

Stable q-ESS states can only develop from a hectic phase when some positions, due to fluctuations, acquire sufficient fitness to be consistent with the q-ESS balance given by equation (13). That is, fluctuations in the weight functions in the hectic phase must allow

$$H_{\text{hectic}} + \frac{\alpha}{C} \geq H_q \quad (16)$$

where $\alpha \in (0, 1)$ describes the width of the distribution of weight functions in the hectic phase, see figure 2, and C determines the width of the distribution of the possible coupling strengths

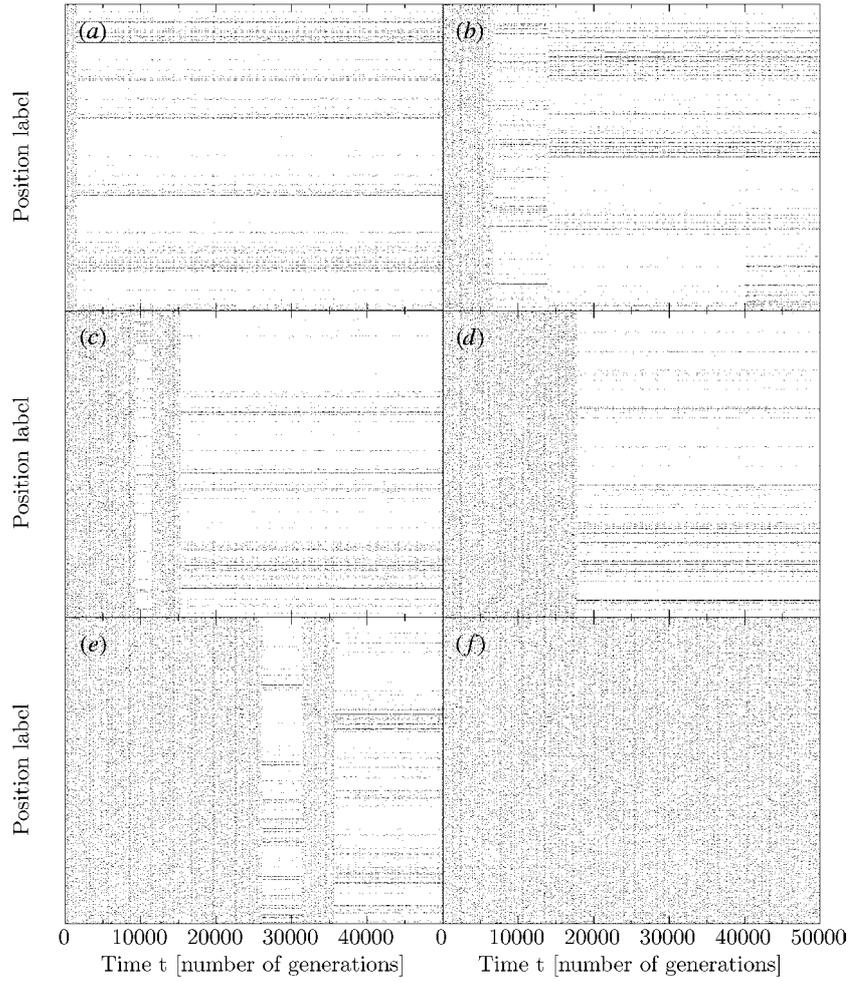


Figure 3. Occupation plots for different values of the mutation rate. The y-axis refers to an arbitrary enumeration of all positions in genotype space. Occupied positions are indicated by a black dot. Results shown are for $p_{kill} = 0.2$, $\mu = 1/1000 \cdot \ln\left(\frac{1-p_{kill}}{p_{kill}}\right)$ and $C = 0.05$. (a) Mutation rate: $p_{mut} = 0.009$. The initial transient is extended. (b) Mutation rate: $p_{mut} = 0.00925$. The initial transient has the same extension of any q-ESS state. (c) Mutation rate: $p_{mut} = 0.0095$. The transition between two q-ESS state are extended. (d) Mutation rate: $p_{mut} = 0.01$. The initial transient is very extended. (e) Mutation rate $p_{mut} = 0.0104$. The initial transient and any transitions are extensively hectic. (f) Mutation rate $p_{mut} = 0.0108$. There is no q-ESS state.

between the individuals. Small C corresponds to the strong coupling regime while large C corresponds to the weak coupling limit. Using equation (14) and equation (15) we obtain

$$\ln\left(\frac{p_{kill}}{1-p_{kill}}\right) + \frac{\alpha}{C} \geq \ln\left(\frac{p_{kill}}{2p_0 - 1 - p_{kill}}\right) \quad (17)$$

which, translated to the mutation rate p_{mut} , becomes

$$p_{mut} \leq 1 - \left[\frac{e^{-\alpha/C}(1-p_{kill}) + 1 + p_{kill}}{2} \right]^{1/L} = p_{mut}^{th} \quad (18)$$

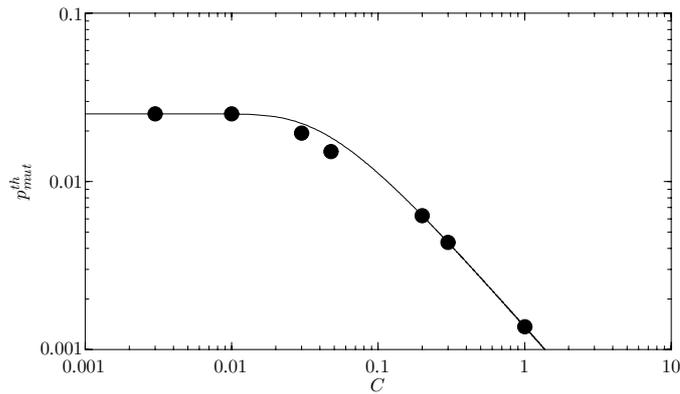


Figure 4. The computational determination of the error threshold. The loss of q-ESS states occurs for mutation rates above the solid circles. The data, compared with the theoretically predicted error threshold p_{mut}^{th} (solid line), indicate a value of $\alpha = 0.07$, see equation (18). The parameters of the simulations are $L = 20$, $\mu = 0.005$ and $p_{kill} = 0.2$.

Equation (18) defines the functional dependency of the error threshold in terms of α , C and p_{kill} . In figure 4 we use α as a fitting parameter and show p_{mut}^{th} as a function of C .

The error threshold has been determined numerically by iterating many simulations with increasing value of the mutation rate for a given C . When no q-ESS emerges, we have reached the error threshold; the lowest p_{mut} for which only a hectic state exists is the estimated value of p_{mut}^{th} .

The numerical results confirm the theoretical predictions given by equation (18) and, qualitatively, are in line with the results of Eigen *et al* [3, 4]. The transition in the tangled nature model appears to be sharp, that is, for values of p_{mut} greater than p_{mut}^{th} q-ESS states are impossible, while for $p_{mut} \leq p_{mut}^{th}$ q-ESS are possible, see figure 3.

Since the factor α represents the width of the distribution of the weight function during a hectic state it is linked to $\mathbf{J} = \{J_{ab}\}$, the set of interactions, and also to μ . This makes it difficult to analytically determine α .

5. Discussion

In the tangled nature model the competition between the organisms is described by the mutual interactions, creating a dynamical rugged fitness landscape where the fitness of a position is determined by the temporal evolution of the ecology. The dynamics, illustrated in [14, 15] selects few heavily occupied positions in genotype space surrounded by other sequences in the immediate vicinity. The organisms in the central positions are only able to reproduce actively. They sustain themselves and all the surrounding ecology. This situation is possible only as long as the mutual interactions are sufficient to counterbalance the dispersive action caused by mutations.

Thus we have derived an interpretation of the tangled nature model as an evolutionary quasi-species model. In the tangled nature model however, the fitness landscape is not fixed. Due to the frequency-dependent fitness landscape, the tangled nature model allows the emergence of multiple coexisting quasi-species during q-ESS states. Also, it should be noted, that in contrast to the model by Eigen *et al* [3, 4], the quasi-species in the tangled nature model are not absolute quantities but may change from one q-ESS to another.

We have discussed and identified the error threshold in the tangled nature model as the mutation rate at which the model is unable to support, over extended periods in time, the occupation of well-defined multiple coexisting genotypes. A formula for the parameter dependence of the error threshold was derived, see equation (18). In particular, the error threshold depends on the genome length as $1/L$, (for large L) which is consistent with the findings in the quasi-species models, see [5, 13]. This result suggests that the mutation rate per base pair itself is subject to selection in a way to make the mutation per base pair decrease with increasing genome length. This is indeed observed in nature.

Acknowledgments

We are extremely grateful to Matt Hall for very helpful discussions. KC gratefully acknowledges the financial support of UK EPSRC through grant no GR/R44683/01.

References

- [1] Carroll S B 2001 The big picture *Nature* **409** 669
- [2] Peliti L 1997 Introduction to the statistical theory of Darwinian evolution *Preprint cond-mat/9712027*
- [3] Eigen M 1971 Selforganization of matter and the evolution of biological macromolecules *Naturwissenschaften* **58** 465–523
- [4] Eigen M, McCaskill J and Schuster P 1988 Molecular quasi-species *J. Phys. Chem.* **92** 6881–91
- [5] Drossel B 2001 Biological evolution and statistical physics *Adv. Phys.* **50** 209–95
- [6] Franz S and Peliti L 1997 Error threshold in simple landscapes *J. Phys. A: Math. Gen.* **30** 4481–7
- [7] Peliti L 2002 Quasispecies evolution in general mean-field landscapes *Europhys. Lett.* **57** 745–51
- [8] Kolár M and Slanina F 2002 How the quasispecies evolution depends on the topology of the genome space *Physica A* **313** 549–68
- [9] Altmeyer S and McCaskill J S 2001 Error threshold for spatially resolved evolution in the quasispecies model *Phys. Rev. Lett.* **86** 5819–22
- [10] Nilsson M and Snoad N 1999 Error thresholds for quasispecies on dynamic fitness landscapes *Phys. Rev. Lett.* **84** 191–4
- [11] Leuthäusser I 1986 An exact correspondence between Eigen's evolution model and a two-dimensional Ising system *J. Chem. Phys.* **84** 1884–5
- [12] Tarazona P 1992 Error thresholds for molecular quasispecies as phase transitions: from simple landscapes to spin-glass models *Phys. Rev. A* **45** 6038–50
- [13] Galluccio S, Graber R and Zhang Y-C 1996 Diffusion on a hypercubic lattice with pinning potential: exact results for the error-catastrophe problem in biological evolution *J. Phys. A: Math. Gen.* **29** L249–L255
- [14] Christensen K, di Collobiano S A, Hall M and Jensen H J 2002 Tangled nature: a model of evolutionary ecology *J. Theor. Biol.* **216** 73–84
- [15] Hall M, Christensen K, di Collobiano S A and Jensen H J 2002 Time dependent extinction rate and species abundance in a tangled-nature model of biological evolution *Phys. Rev. E* **66** 011904
- [16] Hofbauer J and Sigmund K 1998 *Evolutionary Games and Population Dynamics* (Cambridge: Cambridge University Press)