Replication and Mutation on Neutral Networks

By

Christian Reidys [b,c], Christian V. Forst [a,†], and Peter Schuster [a,c,*]

[a] Institut für Theoretische Chemie der Universität, A-1090 Wien, Austria,

[b] Los Alamos National Laboratory, Los Alamos, NM 87 545, U.S.A.,

and

[c] Santa Fe Institute, Santa Fe, NM 87501, U.S.A.

Submitted to Bull. of Math. Biol.

* Address Correspondence to:

Professor Peter Schuster, Institut für Theoretische Chemie der Universität Wien
Währingerstraße 17, A-1090 Wien, Austria

Phone: ++43 1 40480 669, Fax: ++43 1 40480 660

E-Mail: pks @ tbi.univie.ac.at

† Present Address: Theoretical Biophysics Group, Beckman Institute
University of Illinois at Urbana-Champaign, IL 61801, U.S.A.

Keywords.

Error threshold – neutral evolution – neutral network – random graph – RNA secondary strucuture.

Abstract.

Folding of RNA sequences into secondary structures is viewed as a map that assigns a uniquely defined base pairing pattern to every sequence. This mapping is non-invertible since many sequences fold into the same (secondary) structure or shape. The preimages of the map, called **neutral networks**, are uniquely associated with the shapes and *vice versa*. Random graph theory is used to construct networks in sequence space which are appropriate models for neutral networks as they share most of the evolutionarily relevant properties with them.
The theory of molecular quasispecies (in its simplest version) has been applied to replication and mutation on single-peak fitness landscapes. This concept is extended here by considering evolution on degenerate multi-peak landscapes which originate from neutral networks by assuming that one particular shape is fitter than all others. On such a **single-shape landscape** the superior fitness value is assigned to all sequences belonging to the **master shape** (whose associated neutral network comprises all vertices of the corresponding graph). All other shapes are lumped together and their fitness values are averaged in a way that is reminiscent of mean field theory. Replication and mutation on neutral networks are modeled by reformulated phenomenological rate equations as well as by a stochastic birth-and-death model. In analogy to the evolution of quasispecies in sequence space a **phenotypic error threshold** is observed which separates two scenarios: (i) a stationary (fittest) master shape surrounded by closely related shapes and (ii) populations drifting through shape space in a diffusion like process. The error classes of the quasispecies model are replaced by distance classes between the master shape and the other structures.
Analytical results are derived for single-shape landscapes, in particular, simple expressions are obtained for the mean fraction of master shapes in the population and for phenotypic error thresholds. The analytical results are complemented by data obtained from computer simulation of the underlying stochastic processes. The predictions of the phenomenological approach on the single-shape landscape are very well reproduced by replication and mutation kinetics of $\mathrm{tRNA}^{\mathrm{phe}}$. Simulation of the stochastic process resolved in distance classes produces data which are in excellent agreement with the results obtained by the birth-and-death model.

1. Neutral evolution and molecular evolutionary biology

The notion of neutral evolution was coined by Motoo Kimura [27, 28] in order to account for the well established fact that almost all mutations recorded at the level of DNA (RNA or protein) sequences in nature appear to be selectively neutral [29]. Most of the point mutations occur at frequencies which are apparently independent of the morphological changes observed in species. They are used successfully as the basis of a molecular clock measuring time in phylogeny. In essence, the neutral theory reflects relations between genotypes and phenotypes and confirms that a substantial fraction of point mutations does not change fitness. It implicitly assumes that the primary role of selection is to eliminate deleterious variants. A possible existence of neutral variants has indeed been acknowledged by Charles Darwin himself [2]. As a matter of fact, the neutral theory does not (necessarily) contradict Darwin's principle since adaptive mutations might either be rare events which do not show up in statistics or they might occur in the much less well known and phylogenetically younger regulatory proteins that control, for example, cell differentiation and development. Motoo Kimura's view that neutral evolution is an unavoidable byproduct of mutation, caused by the relations between biopolymer sequences, structures and functions, contrasts an older view articulated by Sewall Wright [45, 46] who saw an important role for genetic drift by improving the search capacity of populations in sequence space. Recent studies based on computer simulation of RNA structure optimization [14, 15, 24] favor Wright's rather than Kimura's ideas.

The simplest experimental model system of evolution was discovered by Sol Spiegelman [39] in his test tube assay of RNA replication and mutation. His work was complemented by the seminal paper of Manfred Eigen [5] who presented a theory of molecular evolution based on chemical kinetics. Within this theoretical frame the theory of molecular quasispecies was developed [7, 8]. Evolutionary optimization and adaptation to the environment based on Darwin's principle are not bound to the existence of cellular life. They occur also *in vitro* provided the following minimal requirements are fulfilled:

- a population of objects which are capable of **replication**,
- a replication mechanism that allows for inheritable variation through **mutation**, and
- a reservoir of possible genotypes that is of **combinatorial complexity**.

The third prerequisite is necessary to guarantee sufficient capacity for variation (There are, for example, 4^n different DNA or RNA sequences of chain length n). Basic to the Darwinian scenario is the fundamental dichotomy of evolutionary optimization which states that all variation is done on the genotype through mutations occuring uncorrelated to selection which operates on the phenotypes.

The quasispecies describes the internal population structure of a stationary ensemble of objects replicating independently with a given mutation rate. If only point mutations are considered the error rate can be expressed by a single parameter (p). At some critical error rate (p_{cr}) the quasispecies becomes unstable and the population drifts randomly through sequence space. Quasispecies were studied on various model landscapes including realistic ones derived from RNA folding [12, 13, 14, 15, 24], the simplest being single-peak landscapes with one priviledged genotype having a higher fitness value than all others [31, 40] (This landscape can be understood also as a mean field approximation to the real situation that is obtained by taveraging over the fitness values of all genotypes except the best one). Other more complex and more realistic fitness landscapes have been studied as well. In general, the error threshold phenomenon is retained (for a review see [7]) but there are also somewhat unrealistic cases where the transition to random replication is smooth [44].

This paper is dealing with RNA molecules, the simplest thinkable objects that do indeed fulfill all three requirements listed above. As was already pointed out by

Sol Spiegelman [39] RNA molecules are of particular importance since the whole genotype-phenotype dichotomy is encoded in a single molecule, the genotype being the sequence of nucleotides and the phenotype being the three-dimensional molecular structure with all its properties relevant for differential fitness and selection. Contemporary theoretical biophysics is commonly concerned with structure predictions of biopolymers: Given a sequence; which structure does it form under the specified conditions? This "one sequence-one structure" approach is valuable for many problems. It is, however, neither suitable nor sufficient for modeling evolution. Information is required on the set of structures into which sets of sequences are folded. In RNA evolution the relation between genotypes and phenotypes is tantamount to the mapping of RNA sequences onto molecular structures which are approximated here by coarse grained and simplified versions, the so-called secondary structures (represented by listings of Watson-Crick and **GU** base pairings). Following the ideas of Christian Reidys *et al.* [32] sequence-structure maps can be approximated very well by random graphs. All sequences folding into the same structure form a **neutral network** which is represented by a (random) graph. These networks represent the essential objects in the RNA folding landscape since they constituate the elements of an (uniquely) invertible map from sets of sequences onto structures and *vice versa*. Any evaluation of structures yielding a scalar quantity (fitness being just one of them) induces a **landscape** upon sequence space [35, 36].

Fitness landscapes derived from folding RNA sequences into secondary structures represent the basis for excellent case studies on evolutionary dynamics: optimization processes on neutral networks in sequence space combine the essentials of the quasispecies concept and neutral evolution. Computer simulation of evolutionary optimization and neutral evolution on a landscape derived from $\mathrm{tRNA}^{\mathrm{phe}}$ were reported recently [24]. Here, we present analytical approaches based on phenomenological equations as well as stochastic processes and we test the expressions derived by means of computer simulation data. Evolution in the real world (in cases where it can be appropriately described as optimization of a fitness function) is compromising between adaptive walks on the landscape and random drift. The single-peak landscape model has been transferred to phenotype or shape space: a fittest shape is assumed to be priviledged all others being equal with respect to selection. This approximation allows to derive analytical formulas for error thresholds on neutral networks [34]. The theoretical predictions were subjected to test by computer simulation.

The RNA *in vitro* evolution system is presumably the simplest model of evolutionary adaptation in the Darwinian sense. Experimental data did not only prove the justification of the model but provided also the basis for applied molecular evolution, an already established new area of lively interest in biotechnology. In addition, the quasispecies model has been applied successfully to the evolution of viruses [3, 6]. Populations of RNA viruses share high genetic diversity with those of RNA molecules replicating in test-tubes. Although virus populations live in rapidly varying environments and thus presumably never reach stationarity, the quasispecies concept indeed has been adapted to real systems and provided new insights into virus evolution.

In section 2 we introduce the concept of neutral networks and model them by means of random graph theory. Section 3 introduces and defines the concept of a phenotypic error threshold for RNA secondary structures. Analytical expressions are derived for the single shape landscape and tested by computer simulation. In section 4 these expressions are subjected to test by means of computer simulations of populations on realistic landscapes. One example is replication and mutation of $tRNA^{phe}$.

2. Neutral networks of RNA secondary structures

An RNA secondary structure is a listing of Watson-Crick and **GU** base pairs that can be projected onto a plane without containing knots or pseudoknots. Straightforward application of combinatorics [43] to the elements of RNA secondary structures provide an estimate on the number of acceptable secondary structures[1] formed by the sequences of constant chain length n [37]:

$$S_n \ \approx \ 1.4848 \times n^{-3/2} \, (1.8488)^n \ .$$

This number, although exponentially increasing, is much smaller than the numbers of natural or binary (**GC** or **AU**) RNA sequences which increase with 4^n or 2^n, respectively. The mapping from sequences into (secondary) is there fore highly redundant or many-to-one. Random graph theory applied to hypercubes [33] provides useful tools for an analytical approach to these redundant mappings of RNA molecules [32]. Here we present shortly the results which are important for the forthcoming analysis.

2.1. RNA secondary structures and compatible sequences

Throughout this contribution we shall assume that the chain length of RNA molecules remains unchanged. Then the space of all sequences is the generalized hypercube $\mathcal{Q}_\alpha^n$. The chain length of the polynucleotiedes is n, the size of the alphabet is given by α and the sequences are repesented by the elements of the vertex set, $\mathrm{v}[\mathcal{Q}_\alpha^n]$, of the hypercube. A mapping from sequences into structures is interpreted as the result of a "folding" process that yields the three dimensional structures of the RNA molecules. Predictions of RNA 3D structures from known sequences are still kind of an art and far from being routine [38]. In addition, systematic studies would exceed the capabilities of present day computing systems.

[1] A structure is acceptable when it does not violate obvious steric and energetic criteria. Hairpin loops have to contain at least three unpaired nucleotides and single base pairs are not allowed.

Secondary structures, however, are readily derived by means of efficient computer algorithms [23, 47]. This does not necessarily mean that secondary structures are predicted in agreement with experimental data but there are evolutionarily relevant statistical properties of ensembles of RNA molecules which are reflected correctly by means of the current techniques [41]. The relative simplicity of secondary structures and their accessibilty by fast computer algorithms has invited a great number of analytical and computational investigations.

Secondary structures, in essence, are listing of Watson-Crick and **GU** base pairs which are made precise by the following definition [43]:

> A secondary structure is a vertex-labeled graph on n vertices with an adjencency matrix $A \doteq (a_{i,k})_{1\le i,k\le n}$ fulfilling three conditions:
>
> (1) $a_{i,i+1} = 1$ for $1 \le i \le n-1$,
>
> (2) for each i there is at most one $k \neq i-1, i+1$ with $a_{i,k} = 1$, and
>
> (3) if $a_{i,j} = a_{k,\ell} = 1$ and $i < k < j$ then $i < \ell < j$.
>
> An edge (i,k) with $k \neq i-1, i+1$ is called a bond or a base pair. It is part of the (secondary) structure: $[i,k] \in s$. A vertex i connected only to its neighbors in the backbone chain, $i-1$ and $i+1$, is called unpaired. The numbers of base pairs and unpaired nucleotides in the structure s are denoted by $n_p(s)$ and $n_u(s)$, respectively. The size of the alphabet is denoted by α and the number of classes of base pairs by β.

The chain lenghts of the molecule is $n = n_u(s) + 2n_p(s)$. Condition (3) implies that the secondary structure is a planar graph without knots or pseudoknots.

A pairing rule Π on an alphabet $\mathcal{A}$ is a symmetric relation. For an arbitrary secondary structure we have the pairing rule or set of (non-trivial) contacts

$$\Pi(s) \doteq \big\{[i,k] | a_{i,k} = 1, k \neq i-1, i+1\big\} \ .$$

A sequence or a vertex of the hypercube, $x \in \mathrm{v}[\mathcal{Q}^n_\alpha]$, is compatible with s if and only if holds that $\forall [i,j] \in \Pi(s) : [x_i, x_j] \in \Pi$, which means the nucleotides in the positions i and j are capable of forming a base pair since they are in Π and this is true for all pairs $[i, j] \in \Pi(s)$. The set of all **compatible sequences** is denoted by $\mathbf{C}[s]$. The cardinality of this set, i.e. the number of sequences that are compatible with structure s, is $\alpha^{n_u(s)}\beta^{n_p(s)}$.

Considering a combinatory map, $f_n : \mathcal{Q}_\alpha^n \to \mathcal{S}_n$, we know *a priori* that the vertex set of the preimage $f_n^{-1}(s)$ which consists of all sequences folding into the secondary structure s is contained in the set of compatible sequences. In particular, all neutral neighbors of a sequence x are located in the set $\mathbf{C}[f_n(x)]$. The induced subgraph $\mathcal{Q}_\alpha^n[\mathbf{C}[f_n(x)]]$, however, is not connected – it decomposes into hyperplanes defined by the particular choice of base pair classes.[2] Therefore we introduce the graph $\mathcal{C}[s]$:

> Let s be a secondary structure, then the graph of compatible sequences is $\mathcal{C}[s] \cong \mathcal{Q}_\alpha^{n_u(s)} \times \mathcal{Q}_\beta^{n_p(s)}$. Obviously $\mathcal{C}[s]$ has the vertex set $\mathbf{C}[s]$ and by definition of the product of graphs, two sequences $x, y \in \mathbf{C}[s]$ are neighbors, if they differ either
>
> - in a single position i which is unpaired in s, or
> - in two positions i and j which form a base pair $[i, j] \in s$.

Two graphs of compatible sequences $\mathcal{C}[s], \mathcal{C}[s']$ are isomorphic as graphs if and only if both have the same number of unpaired and paired nucleotides. Therefore two different secondary structures $s, s' \in \mathcal{S}_n$ can lead to one and the same graph of compatible sequences.

2.2. Neutral networks as random graphs

The combinatory map is a (non-invertible) map from sequence space into shape space, $f_n : \mathcal{Q}_\alpha^n \to \mathcal{S}_n$, defines the neutral network $\Gamma_n[s]$ for a given secondary

[2] Accessibility through single point mutations leads to a partitioning of base pairs into two disconnected groups: $[\mathbf{AU}] \leftrightarrow [\mathbf{GU}] \leftrightarrow [\mathbf{GC}]$ and $[\mathbf{CG}] \leftrightarrow [\mathbf{UG}] \leftrightarrow [\mathbf{UA}]$. We note that the two triples of base pairs are characterized by constant positions of pyrimidines and purines, $[\mathbf{RY}]$ and $[\mathbf{YR}]$, respectively.

structure s:

$$\Gamma_n[s] \doteq \mathcal{C}[s][f_n^{-1}(s)].$$

Neutral networks were modeled by random graphs in sequence space in order to provide a theoretical concept that allows to derive analytical results which can be used as an exact reference for networks derived from real structures [16, 32]. In order to construct the neutral network of a given structure we consider first the set of compatible sequences. From this set sequences are chosen at random but unpaired nucleotides and base pairs are distinguished. Allowed moves within the set of compatible sequences are thus base exchanges for unpaired nucleotides and pair exchanges for base pairs (see previous section). Each sequence, x_i, has a certain number of neutral neighbors, say U_i, in the single base exchange neighborhood and in the single base pair exchange neighborhood, say P_i. The fractions of neutral neighbors is given by $(\lambda_u)_i = U_i \,/\, (\alpha-1)\, n_u$ and $(\lambda_p)_i = P_i \,/\, (\beta-1)\, n_p$, respectively. Averaging over all sequences of the neutral net yields two neutrality parameters

$$\lambda_u = \frac{1}{|\Gamma_n[s]|} \sum_{i\in\Gamma_n[s]} (\lambda_u)_i \quad \text{and} \quad \lambda_p = \frac{1}{|\Gamma_n[s]|} \sum_{i\in\Gamma_n[s]} (\lambda_p)_i \ .$$

There is no *a priori* reason why the probability to create a neutral neighbor in sequence space should be the same for a single base exchange and an exchange of a base pair and, in general, λ_u will be different from λ_p.

The two classes of mutations are assumed to be fully correlated with respect to their consequences in the space of structures. The regularities in shape space suggest the construction of a probability space formed by random induced subgraphs of the hypercube $\mathcal{Q}_\alpha^n$.

Each subset $X \subset \mathrm{v}[\mathcal{Q}_\alpha^n]$ induces a subgraph $\mathcal{Q}_\alpha^n[X]$ establishing a one-to-one correspondence between $X \subset \mathrm{v}[\mathcal{Q}_\alpha^n]$ and $\mathcal{Q}_\alpha^n[X]$. We denote the set of all induced subgraphs of $\mathcal{Q}_\alpha^n$ by $\mathcal{G}(\mathcal{Q}_\alpha^n)$ and suppose $0 \le \lambda \le 1$. In order to derive a probability measure $\boldsymbol{\mu}_\lambda$ we set for $\Gamma_n \in \mathcal{G}(\mathcal{Q}_\alpha^n)$

$$\boldsymbol{\mu}_\lambda\{\Gamma_n\} \doteq \lambda^{|\mathrm{v}[\Gamma_n]|}(1-\lambda)^{\alpha^n - |\mathrm{v}[\Gamma_n]|}.$$

Random graph theory states the existence of a threshold value λ^* for the density and the connectivity property of random induced subgraphs on generalized hypercubes. A subgraph $\Gamma_n < \mathcal{Q}_\alpha^n$ is called **dense** in $\mathcal{Q}_\alpha^n$ if and only if $\overline{\mathrm{v}[\Gamma_n]} = \mathrm{v}[\mathcal{Q}_\alpha^n]$. Proofs for the two main results are found in [32].

Given a threshold value $\lambda^* \doteq 1 - \sqrt[\alpha-1]{\alpha^{-1}}$, then we obtain

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n \text{ is } \mathbf{dense}\} = \begin{cases} 1 & \text{for} \quad \lambda > 1 - \sqrt[\alpha-1]{\alpha^{-1}} \\ 0 & \text{for} \quad \lambda < 1 - \sqrt[\alpha-1]{\alpha^{-1}}. \end{cases}$$

Let $(\mathcal{Q}_\alpha^n)$ be a sequence of generalized hypercubes and $\Gamma_n < \mathcal{Q}_\alpha^n$ random induced subgraphs. Then

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n \text{ is } \mathbf{connected}\} = \begin{cases} 1 & \text{for} \quad \lambda > 1 - \sqrt[\alpha-1]{\alpha^{-1}} \\ 0 & \text{for} \quad \lambda < 1 - \sqrt[\alpha-1]{\alpha^{-1}}. \end{cases} \tag{1}$$

$$\begin{array}{ccccc} \Gamma_{n_u} & \overset{\text{in}}{\hookrightarrow} & \Gamma_{n_u} \times \Gamma_{n_p} & \overset{\text{in}}{\hookleftarrow} & \Gamma_{n_p} \\ & & \text{in} \downarrow & & \\ & & \mathcal{Q}_\alpha^{n_u} \times \mathcal{Q}_\beta^{n_p} & & \end{array}$$

Diagram 1: Relations between the subgraphs referring to unpaired and paired parts of RNA secondary structures and their embedding in the space of compatible sequences.

Taking the graph product of the two random induced subgraphs corresponding to the unpaired and paired parts of the secondary structure s as illustrated in diagram 1 we obtain a probability space of random induced subgraphs of $\mathcal{Q}_\alpha^{n_u} \times \mathcal{Q}_\beta^{n_p}$ as follows:

Let $\Gamma_{n_u} \in \mathcal{G}(\mathcal{Q}_\alpha^{n_u})$ and $\Gamma_{n_p} \in \mathcal{G}(\mathcal{Q}_\beta^{n_p})$ be random subgraphs as introduced in the above model. We set $\Gamma_n[s] \doteq \Gamma_{n_u} \times \Gamma_{n_p}$ and

$$\boldsymbol{\mu}_{\lambda_u,\lambda_p}(\Gamma_n[s]) \doteq \boldsymbol{\mu}_{n_u,\lambda_u}(\Gamma_{n_u}) \otimes \boldsymbol{\mu}_{n_p,\lambda_p}(\Gamma_{n_p}) .$$

Then $\boldsymbol{\mu}_{\lambda_u,\lambda_p}$ is a probability measure and $\Gamma_n[s] < \mathcal{C}[s]$.

Random graph theory provides simple analytical expressions that allow to predict generic structures of neutral networks. These predictions are fulfilled very well unless certain structural features require modifications [21, 22]. Most relevant for replication-mutation dynamics is the fact that the neutral networks of common structures fulfill almost always the connectivity condition [20].

3. Phenotypic error thresholds

In order to be able to analyse replication dynamics on neutral networks we need to assign fitness values to phenotypes. Realistic fitness landscapes are complicated objects but they can be approximated well by simplifications (see section 4.1. In order to study generic features of error propagation in replication it is sufficient to use a single-peak landscape in shape space which distinguishes a **master shape** from all others by assigning higher fitness to it. Then the neutral network of the master shape $\Gamma[m]$ induces the fitness landscape being a mapping $f_{\Gamma[m]} : \mathcal{Q}_\alpha^n \to \mathbb{R}_+$:

$$f_{\Gamma[m]}(v) \doteq \begin{cases} \sigma > 1 & \text{iff} \quad v \in \mathrm{v}\big[\Gamma[m]\big] \\ 1 & \text{iff} \quad v \in \mathrm{v}\big[\mathcal{Q}_\alpha^n\big] \setminus \mathrm{v}\big[\Gamma[m]\big] . \end{cases}$$

The single-peak landscape reminds of the mean field approximation often applied in physics: The fitness of the master species is seen relative to an averaged fitness of the mutant distributions (see section 3.1).

The error threshold will be derived first by means of a phenomenological approach based on the kinetic differential equation for replication and mutation [5], then a stochastic replication-deletion process will be used in the derivation. In section 4 numerical values for the threshold are compared with the results from computer simulations [18, 19].

3.1. Genotypic and phenotypic error thresholds

Error thresholds in replication-mutation equations have been studied extensively in systems lacking selective neutrality [5, 7, 8]. We repeat shortly the basic results of the quasispecies concept and extend it then to systems with selective neutrality expressed by the average fraction of neutral neighbors (λ).[3] The time dependence of the frequency of genotype x_i in the population, ξ_i, is determined by the kinetic equation

$$\dot{\xi}_i = \Big(a_i Q_{ii} - d_i - \bar{E}\Big)\xi_i + \sum_{j \neq i} a_j Q_{ji} \xi_j \ , \quad i,j = 1,\ldots,N \ .$$

Genotype frequencies are normalized: $\sum_{i=1}^{N} \xi_i = 1$. The rate constants for replication and degradation of the molecular species x_i are a_i and d_i, respectively. Correct replication and mutations are understood as parallel reaction channels whose relative frequencies are contained in the matrix $Q \doteq \{Q_{ij};\, j = 1,\ldots,N\}$: The fraction of (error) copies of template x_i leading to genotype x_j is given by Q_{ij}. Combinations of rate constants and mutation rates turned out to be useful: the excess productivity $e_i = a_i - d_i$, its mean value averaged over the entire population $\bar{E} = \sum_{i=1}^{N} e_i \xi_i$, and the value matrix $W \doteq \{w_{ij} = a_i Q_{ij} - d_i \delta_{ij};\, i,j = 1,\ldots,N\}$ with $\delta_{ij} = 1$ or 0 for $i = j$ and $i \neq j$, respectively. The diagonal elements of the value matrix are called selective values. The selective value of a genotype is

[3] The fraction of neutral neighbors, λ_i, refers here to the total one-error neighborhood of the sequence x_i averaged over the network $\Gamma[s]$. No partitioning into unpaired and paired regions of s is assumed.

tantamount to its fitness in case of constant total number of genotypes and vanishing mutational backflow ($\sum_{j \neq i} a_j Q_{ji} \xi_j$). Accordingly, the genotype with optimal selective value, x_m,

$$w_{mm} \;=\; \max(w_{ii} | i = 1, \ldots, N) \;,$$

dominates a populations when it is at selection equilibrium and hence is called the **master** sequence. We assume $d_1 = d_2 = \ldots = d_N = d$ in order to simplify the forthcoming analysis without losing generality. The stationary value of the frequency of the master genotype ($\bar{\xi}_m$) is readily derived from the conditions $\dot{\xi}_i = 0;\;\; i + 1, \ldots, N$. Indeed, the stationary distribution of genotype frequencies can be computed exactly as the solution of an eigenvalue problem [25, 42].

In correspondence with the single peak landscape master genotypes (x_m) are distinguished from all other genotypes ($x_j;\; j \neq m$ which are lumped together to form a mutant cloud with $\sum_{j \neq m} \xi_j = 1 - \xi_m$). Reminiscent of mean field approximations often used in physics we introduce a "mean except the master" for replication rate constants and excess productivities, $\bar{a} = \sum_{j \neq m} a_j \xi_j \;/\; (1 - \xi_m)$ and $\bar{e} = \sum_{j \neq m} (a_j - d)\, \xi_j \;/\; (1 - \xi_m)$, respectively, and define a **superiority** of the master sequence, $\sigma_m = a_m \;/\; \bar{a}$. Neglecting the mutational backflow we obtain for the stationary concentration of the master sequence:

$$\bar{\xi}_m = \frac{w_m - \bar{e}}{e_m - \bar{e}_m} = \frac{a_m Q_{mm} - \bar{a}}{a_m - \bar{a}} = \frac{\sigma_m Q_{mm} - 1}{\sigma_m - 1} \;.$$

A simple but, nevertheless, fairly accurate estimate called the zero-th order approximation (see p.177 in [7]) to the localization threshold of molecular quasispecies is obtained by computing the replication accuracy (Q_{mm}) at which the stationary concentration of the master genotype vanishes:

$$Q_{mm}(\bar{\xi}_m = 0) = Q_{min} = \sigma^{-1} \;.$$

Application of the uniform error rate model [8][4] yields for the genotypic error threshold:

$$q_{min} \;=\; 1 - p_{max} \;=\; \sigma^{-1/n} \;, \tag{2}$$

where $p = 1 - q$ is the mutation rate per (nucleotide) site. In case of selective neutrality ($a_m \to \bar{a}$) the error threshold converges to the limit of absolute replication accuracy ($q_{min} \to 1$). Thus, the model in its original form cannot be used to describe evolutionary stability of phenotypes in the selectively neutral case.

In order to deal with the neutral case we modify the kinetic equations in straightforward way. Genotypes are ordered with respect to non-increasing fitness. We assume that the first k different genotypes have maximal fitness, $w_1 = w_2 = \ldots = w_k = \tilde{w}_m = w_{max}$, as well as identical replication and degradation rate constants $a_1 = a_2 = \ldots = a_k = \tilde{a}_m$ and $d_1 = d_2 = \ldots = d_k = \tilde{d}_m$. In addition, we define new variables, η_j ($j = 1, \ldots, M$), that lump together all genotypes folding into the same phenotype:

$$\eta_j \;=\; \sum_{i\,=\,k_{j-1}+1}^{k_j} \xi_i \;\text{ with }\; \sum_{j=1}^{M} \eta_j = \sum_{i=1}^{N} \xi_i = 1 \;,$$

Without a loss of generality we characterize the master phenoptype by "m", $\eta_m = \sum_{i=1}^{k} \xi_i$, and obtain the following kinetic differential equation for the set of sequences forming the neutral network of the phenotype with maximal fitness:

$$\dot{\eta}_m \;=\; \sum_{i=1}^{k} \dot{\xi}_i \;=\; \eta_m \left(\tilde{a}_m Q_{kk} - \tilde{d}_m - \sum_{j=1}^{M} \tilde{e}_j \eta_j \right) \;+\; \sum_{i=1}^{k} \sum_{j \neq i} a_j Q_{ji} \xi_j \;.$$

The mean excess productivity of the population is, of course, independent of the choice of variables:

$$\bar{E} \;=\; \sum_{j=1}^{M} \tilde{e}_j \eta_j \;=\; \sum_{i=1}^{N} e_i \xi_i \;.$$

[4]The uniform error rate model assumes that the single digit replication accuracy (q) is independent of nucleotide and site. Hence, we obtain $Q_{ii} = q^n (i=1,\ldots,N)$ for polynucleotides of chain length n.

In order to derive a suitable expression for a phenotypic error threshold we split the mutational backflow into two contributions, (i) mutational backflow on the neutral network and (ii) mutational backflow from genotypes not on the network:

$$\sum_{i=1}^{k}\sum_{j\neq i} a_j Q_{ji}\xi_j \;=\; \left\{\tilde{a}_m \sum_{i=1}^{k}\sum_{j=1,j\neq i}^{k} Q_{ji}\xi_j\right\} + \left\{\sum_{i=1}^{k}\sum_{j=k+1}^{N} a_j Q_{ji}\xi_j\right\} \;.$$

The first term is readily computed under two assumptions which are approximately fulfilled: (i) the fraction of selectively neutral neighbors (λ_m) is constant and (ii) mutation rates are equal ($Q_{ji} = \bar{Q}_j; i,j = 1,\ldots,k; i \neq j$) on the master network:

$$\begin{aligned}\sum_{i=1}^{k}\sum_{j=1,j\neq i}^{k} Q_{ji}\xi_j \;&\approx\; \frac{\lambda_m(1-Q_{mm})}{k-1}\sum_{i=1}^{k}\sum_{j=1,j\neq i}^{k}\xi_j \;= \\ &=\; \frac{\lambda_m(1-Q_{mm})}{k-1}\sum_{j=1,j\neq i}^{k}\sum_{i=1}^{k}\xi_j \;=\; \lambda_m(1-Q_{mm})\,\eta_m \;.\end{aligned}$$

Mutational backflow from other networks ($\eta_j, j \neq m$) need not be evaluated explicitly since it has also been neglected in the derivation of the genotypic error threshold. The kinetic equation for the master phenotype can now be rewritten:

$$\dot{\eta}_m \;=\; \left(\tilde{a}_m\tilde{Q}_{mm} - \tilde{d}_m - \bar{E}\right)\eta_m \;+\; \text{Mutational Backflow}$$

They are identical with those in the variables expressing genotype concentrations except the use of an effective replication accuracy of

$$\tilde{Q}_{mm} \;=\; Q_{mm} + \lambda_m(1-Q_{mm}) \;=\; q^n\left(\Phi(q)\,\lambda_m + 1\right), \text{ with } \Phi(q) = \left(\frac{1}{q^n} - 1\right). \quad (3)$$

Neglecting mutational backflow from non-master phenotypes we finally find in complete analogy with the derivation of the genotypic error threshold

$$\tilde{Q}_{min} \;=\; \left\{Q_{mm} + \lambda_m(1-Q_{mm})\right\}_{min} \;=\; \sigma_m^{-1}$$

where σ_m is the superiority of the "master phenotype". Introducing the uniform error model we obtain by neglect of mutational backflow for the stationary frequency of master phenotypes:

$$\bar{\eta}_m(p) \;=\; \frac{\tilde{Q}_{mm}(p)\,\sigma_m \;-\; 1}{\sigma_m - 1} \;=\; \frac{(1-p)^n\,\sigma_m\,(1-\lambda_m) \;+\; \sigma_m\lambda_m \;-\; 1}{\sigma_m - 1} \;. \quad (4)$$

Eventually we find for the phenotypic error threshold by applying the "zero-th order approximation", $\bar{\eta}_m = 0$:

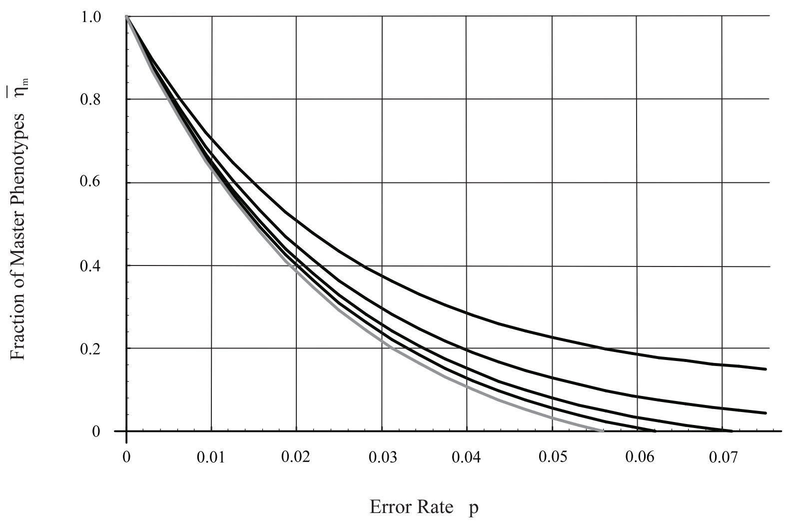
$$q_{min} = (1 - p_{max}) = \left(\frac{1 - \lambda_m \sigma_m}{(1 - \lambda_m)\, \sigma_m} \right)^{1/n} . \tag{5}$$

The function $q = q_{min}(n, \lambda_m, \sigma_m)$ is illustrated in figure 1. The limits are easily visualized: (i) the phenotypic error threshold converges to the genotypic value $q_{min} = \sigma_m^{-1/n}$ in the limit $\lambda_m \to 0$ and (ii) the minimal replication accuracy q_{min} approaches zero in the limit $\lambda_m \to \sigma_m^{-1}$. The second case implies that single digit accuracy plays no role in case the degree of neutrality exceeds the reciprocal value of the superiority.

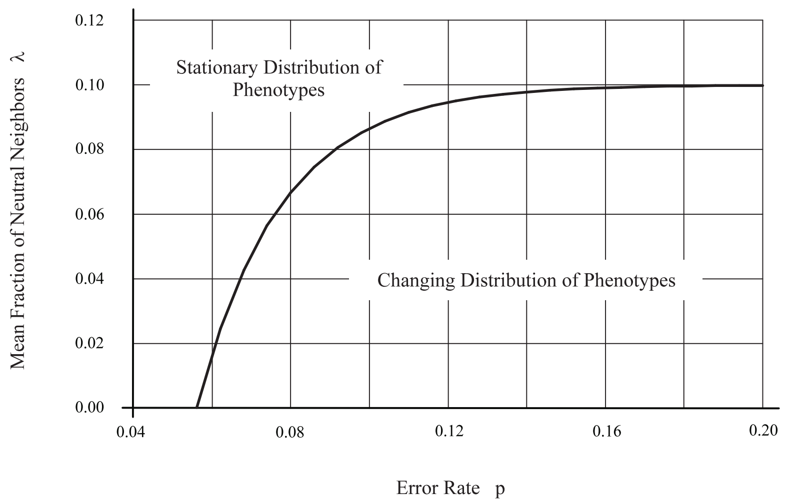
Recapitulating the results on stationary distributions of phenotypes derived in this section we state that selective neutrality allows to tolerate more replication errors than in the non-neutral case. We are dealing with a distribution of changing genotypes corresponding to a population which drifts randomly [24] on the neutral network corresponding to the fittest or master phenotype. In this drift the master phenotype is conserved as long as the replication accuracy is above a critical minimal value, q_{min}. In case the accuracy falls below this critical value the populations drifts through sequence space and through shape space and no more stationarity, neither with genotypes nor with phenotypes, is observed. It is particularly interesting to note that there is a degree of neutrality related to the superiority of the master phenotype ($\lambda = \sigma^{-1}$) above which the error rate does not matter. In other words, the master phenotype will never be lost when the degree of neutrality increases the inverse superiority.

3.2. Replication-Deletion Process

Populations on graphs and their temporal changes are readily described by a general formalism which we apply here for consistence with the random graph model



A



B

Figure 1: The phenotypic error threshold from phenomenological equations. The upper part (**A**) shows the stationary frequency of master phenotypes, $\bar{\eta}_m$, as a function of the error rate p. The parameter values are: n=40, σ_m=10 and λ_m=0 (gray),0.025,0.050, 0.1 and 0.2. The frequency $\bar{\eta}_m(p)$ shows roughly exponential decrease until it becomes very small near the error threshold p_{max}. Exceptions are the two highest curves where the degree of neutrality is equal or larger than the reciprocal superiority ($\lambda_m \geq \sigma_m^{-1}$) and hence the master phenotype is never lost from the population. In the lower part (**B**) we show the error threshold $p_{max}=1-q_{min}$ as a function of the degree of neutrality of the master phenotype, λ_m for n=40 and σ_m=10. It is interesting to note that q_{min} approaches zero when λ_m becomes σ_m^{-1} in agreement with plot **A** showing that errors do no lead to a loss of the master phenotype.

of neutral networks. A population $\mathbf{V}$ is a (finite) family of vertices on the hypercube $\{ v_i \,|\, i \in \mathbb{N}_N \} \subset \mathrm{v}[\mathcal{Q}_\alpha^n]$ where $\mathbb{N}_N$ stands for the natural numbers up to the population size N ($\mathbb{N}_N \equiv \{1, 2, \ldots, N\}$)[5]. In the theory of point processes such a family may be identified by an integer valued measure:

$$\mathbf{V} = (v_i \,|\, i \in \mathbb{N}_N) \quad \longleftrightarrow \quad \phi \doteq \sum_{i=1}^{N} g_{v_i}, \text{ where } g_{v_i}(v) \doteq \begin{cases} 1 & \text{for } v = v_i\,, \\ 0 & \text{for } v \neq v_i\,. \end{cases} \tag{6}$$

This measure contains the information on the numbers of individual genotypes present in the population.

The replication-deletion process is known in mathematical biology as Moran model [30] and consists of two strongly coupled random events: An arbitrarily chosen member of the population is subjected to replication whereas another randomly chosen member is deleted. The population size is constant and strictly equals N between such pairs of coupled events. In our notation a replication-deletion process is a mapping from the family $(v_i \,|\, i \in \mathbb{N}_N)$ into the family $(v_i' \,|\, i \in \mathbb{N}_N)$: An ordered pair (v_l, v_k) is drawn from the population $\mathbf{V}$. The first vertex (v_l) chosen with uniform probability $\sigma\,\ell\,\big/\big((N-\ell)+\sigma\ell\big)$ from the ℓ elements located at the neutral net $\Gamma_n[s]$ – formally expressed by $(v_i \,|\, i \in \mathbb{N}_N$ and $\mathrm{res}_{\Gamma_n[s]}\Phi(v_i) \geq 1)$ – and with uniform probability $(N-\ell)\,\big/\big((N-\ell)+\sigma\ell\big)$ from the remaining $N-\ell$ elements. The number of elements on the neutral net is consistently expressed by: $\ell = \mathrm{res}_{\mathrm{v}[\Gamma_n[s]]}\phi(\mathrm{v}\big[\Gamma_n[s]\big])$. The second vertex (v_k) of the ordered pair is chosen with uniform probability $1/(N-1)$ from all vertices except the first one, $(v_i \neq v_l \,|\, i \in \mathbb{N}_N)$. Two events introduce the changes in the population:

(i) The vertex $v_l = (b_1, ..., b_n)$ is mapped randomly into $v^* = (b_1', ..., b_n')$ by mapping each coordinate b_i with uniform probability p into $b_i' \neq b_i$ $(b, b' \in \mathcal{A})$ and leaving it unchanged otherwise, and

(ii) the vertex (v_k) is deleted.

[5] For formal reasons which will become clear when we define the replication-deletion process we do not consider single vertices as populations: $N \geq 2$.

The random mapping (i), $v_l \mapsto v^*$, is a replication event called error-free copying if $v^* \equiv v_l$ – which occurs with probability $(1-p)^n$ – and mutation otherwise. The complete mapping (i) and (ii), $(v_l, v_k) \mapsto (v_l, v^*)$, is an elementary act of the replication-deletion process.

The individual (complete) mappings are considered as independent events and the time intervals $\hat{T}$ between consecutive mappings are exponentially distributed:

$$\boldsymbol{\mu}\{\hat{T} \leq t\} = e^{-[(N-\ell)+\sigma\ell]\,t}\,,$$

on a time axis which is scaled by the mean fitness of the population.

Accordingly, the replication-deletion process is a stochastic process $(\hat{X}_t)_t$ in continuous time with integer values in

$$M_N(\mathrm{v}[\mathcal{Q}_\alpha^n]) \doteq \{\phi \mid \phi \text{ is an integer valued measure on } \mathrm{v}[\mathcal{Q}_\alpha^n] \text{ and } \phi(\mathrm{v}[\mathcal{Q}_\alpha^n]) = N\,\}\,.$$

As in the previous section 3.1 we lump together all vertices on the neutral network $(\mathrm{v}\big[\Gamma_n[s]\big])$ and all other vertices $(\mathrm{v}[\mathcal{Q}_\alpha^n]\backslash\mathrm{v}\big[\Gamma_n[s]\big])$ and obtain a simplified stochastic process on the natural numbers,

$$(\hat{Y}_t)_t \doteq \left(\left|\hat{X}_t(\mathrm{v}\big[\Gamma_n[s]\big])\right|\right)_t\,, \tag{7}$$

which describes the evolution of the number of elements in the population that are located on the neutral network. As in the kinetic differential equation the population is split into two parts. Precisely, the mapping $f_{\Gamma_n[s]}$ induces a bipartition of the population $\mathbf{V}$ on $\mathcal{Q}_\alpha^n$ according to equation (7):

$$\begin{aligned} \mathbf{V}_\mu &\doteq \{v \in \mathbf{V} \mid v \in \mathrm{v}\big[\Gamma_n[s]\big]\} \quad &\longleftrightarrow& \quad \mathrm{res}_{\mathrm{v}[\Gamma_n[s]]}\phi \\ \mathbf{V}_\nu &\doteq \{v \in \mathbf{V} \mid v \notin \mathrm{v}\big[\Gamma_n[s]\big]\} \quad &\longleftrightarrow& \quad \mathrm{res}_{\mathrm{v}[\mathcal{Q}_\alpha^n]\backslash\mathrm{v}[\Gamma_n[s]]}\phi\,, \end{aligned}$$

where $\mathbf{V} = \mathbf{V}_\mu \dot{\cup}\, \mathbf{V}_\nu$. The elements of $\mathbf{V}_\mu$ are called master phenotypes and those of $\mathbf{V}_\nu$ are the mutant phenotypes.

The analysis of the stochastic process $\hat{Y}_t$ aims at a computation of the distribution function

$$P_\ell(t) \;=\; \mathrm{Prob}\,\{\hat{Y}_t = \ell\}\,; \quad \ell = 0, 1, \ldots, N \;,$$

whose time dependence is commonly described by a master equation (see [17], for example):

$$\frac{\mathrm{d}P_\ell}{\mathrm{dt}} \;=\; \boldsymbol{P}_{\ell+1,\ell}\,P_{\ell+1}(t) \;+\; \boldsymbol{P}_{\ell-1,\ell}\,P_{\ell-1}(t) \;-\; (\boldsymbol{P}_{\ell,\ell+1} + \boldsymbol{P}_{\ell,\ell-1})\;P_\ell(t) \;.$$

Here we made implicitly the common assumption for chemical reactions that all transitions changing ℓ by more than one step are of order $O(\Delta t)$ and thus vanish in the limit $\lim \Delta t \to 0$.

In the forthcoming section 3.3 we shall study the replication-deletion process on a single-shapr fitness landscape by means of an analytical model based on a birth-and-death process and in section 4.2 we present the results of computer simulations using the algorithm introduced by Gillespie [18, 19].

3.3. Birth-and-death model

A birth-and-death process is a Markovian process in continuous time which is completely described by its transition rates, $\boldsymbol{P}_{\ell,\ell+1}$, $\boldsymbol{P}_{\ell,\ell}$ and $\boldsymbol{P}_{\ell,\ell-1}$. We are dealing with a finite population of constant size N which leads to a birth-and-death process that converges to a stationary distribution (Appendix). It is useful to transform the time axis such that the probability for a replication-deletion event becomes independent of the mean of the population. This "reactor time" $\tilde{t}$ is defined by

$$d\tilde{t} \;\doteq\; dt\,\frac{N}{(\sigma-1)\ell \,+\, N} \;=\; dt\,\tau(\ell) \;.$$

Accordingly, we obtain for the birth-and-death rates in the scaled reactor time

$$\mathbf{P}_{\ell,\ell+1} = \frac{(N-\ell)}{N(N-1)}\,\left[\,\sigma\ell\,\bar{W}_{\mu,\mu} + (N-1-\ell)\bar{W}_{\nu,\mu}\right] \text{ for } 0 \le \ell \le N-1\,,$$

$$\mathbf{P}_{\ell,\ell-1} = \frac{\ell}{N(N-1)}\,\left[\,\sigma(\ell-1)\,\bar{W}_{\mu,\nu} + (N-\ell)\,\bar{W}_{\nu,\nu}\right] \text{ for } 1 \le \ell \le N\,, \text{ and}$$

$$\mathbf{P}_{\ell,\ell} = 1 \,-\, \mathbf{P}_{\ell,\ell+1} \,-\, \mathbf{P}_{\ell,\ell-1} \;,$$

and $\mathbf{P}_{\ell,k} = 0$ otherwise.

The mean values, $\bar{W}_{\cdot,\cdot}$, represent averages of the conditional probabilities, $W_{\cdot,\cdot}(v)$, over the whole neutral network $\Gamma_n[s]$. They refer to the mappings discussed for the replication-deletion process (see section 3.2). In particular,

(i) $W_{\mu,\mu}(v)$ is the conditional probability of a mapping of $v \in \Gamma_n[s]$ into $\Gamma_n[s]$: a master vertex maps into a master vertex,

(ii) $W_{\mu,\nu}(v)$ is the conditional probability of a mapping from $v \in \Gamma_n[s]$ into the set $\mathcal{Q}_\alpha^n \setminus \Gamma_n[s]$: a master vertex maps into a vertex of a mutant phenotype,

(iii) $W_{\nu,\mu}(u)$ is the conditional probability of a mapping from $u \in \mathcal{Q}_\alpha^n \setminus \Gamma_n[s]$ onto the neutral network $\Gamma_n[s]$: a mutant phenotype vertex maps into a master vertex, and

(iv) $W_{\nu,\nu}(u)$ is the conditional probability of a particular mapping within the set $\mathcal{Q}_\alpha^n \setminus \Gamma_n[s]$: a mutant phenotype vertex maps into a mutant phenotype vertex.

By definition we have $W_{\mu,\mu}(v) + W_{\mu,\nu}(v) = 1$ and $W_{\nu,\mu}(u) + W_{\nu,\nu}(u) = 1$, and the same is, of course, true for the corresponding mean values. It is thus sufficient to compute the two probabilities $\bar{W}_{\mu,\mu}$ and $\bar{W}_{\nu,\mu}$ for a full description of the process.

3.3.1. Transition probabilities

By construction the neutral network $\Gamma_n[s]$ has binomially distributed vertex degrees since compatible adjacent vertices are selected with independent probabilities, λ_u and λ_p, respectively. Hence, the probabilities refer also to binomially distributed random variables. Application of the birth-and-death process to the stochastic dynamics on the hypercube thus is an approximation which treats the neutral network $\Gamma_n[s]$ as a regular graph. The transition probability $\bar{W}_{\mu,\mu}$ on the product space $\mathcal{Q}_\alpha^{n_u} \times \mathcal{Q}_\beta^{n_p}$ is readily computed (see appendix):

$$\bar{W}_{\mu,\mu} \;=\; (1-p)^n \left(\Phi(p)\,\lambda_u + 1\right) \left(\Psi(p)\,\lambda_p + 1\right) \,, \tag{8}$$

with the two functions $\Phi(p)$ and $\Psi(p)$ expressing the dependence of the coefficients of λ_u and λ_p on the replication accuracy p:

$$\Phi(p) \doteq \frac{1}{(1-p)^{n_u}} - 1 \quad \text{and} \quad \Psi(p) \doteq \left(\frac{p^2}{(\beta-1)(1-p)^2} + 1\right)^{n_p} - 1 \ .$$

Computing of the remaining transition probability requires the notion of **compatibility classes**. For a vertex $v \in \mathrm{v}[\mathcal{Q}_\alpha^n]$ the incompatibility distance with respect to $\Gamma_n[s]$ is

$$d(\Gamma_n[s], v) \doteq |\, \{[v_i, v_k] \,|\, [v_i, v_k] \notin \Pi \wedge [i,k] \in \Pi(s)\} \,|$$

and the i-th compatibility class, $\mathbf{C}_i[s]$, is defined by

$$\mathbf{C}_i[s] \overset{\text{def}}{=\!=} \{v \in \mathrm{v}[\mathcal{Q}_\alpha^n] \setminus \mathrm{v}[\Gamma_n[s]] \,|\, d(\Gamma_n[s], v) = i\} \quad \forall i = 0, \ldots, n_p.$$

The class with $i = 0$ is the class of compatible sequences that do not form the master phenotype. The other classes with $i = 1, 2, \ldots, n_p$ are incompatible with the master phenotype in one, two, $\ldots, n_p$ base pairs, respectively. Next we assume that the mutant elements of the population are uniformly distributed in the corresponding compatiblility classes:

$$\boldsymbol{\mu}\{\phi(v) = k\} \quad \text{is independent of} \quad v \in \mathrm{v}[\mathcal{Q}_\alpha^n] \setminus \mathrm{v}[\Gamma_n[s]] \quad \text{for} \ \ 0 \le k \le N \ .$$

Furthermore the computation of the expectation values is simplified by the assumption of an alphabet in which the number of bases equals the number of base pairs: $\alpha = \beta$. In particular, we find for binary alphabets ($\alpha = \beta = 2$):

$$|\mathbf{C}_i[s]| = \begin{cases} \binom{n_p}{i} 2^{n_p + n_u} & \text{for } 1 \le i \le n_p \\ 2^{n_u + n_p} - |\Gamma_n[s]| & \text{for } i = 0 \end{cases} ,$$

and obtain for the expectation value:

$$\bar{W}_{\nu,\mu} = \sum_{h=1}^{n} \lambda_p \, p^h (1-p)^{n-h} \sum_{\ell=0}^{h} \binom{n_u}{h-\ell} \lambda_u^\delta \sum_{i=0}^{\ell} \frac{N_i}{N-\ell} \, 2^i \binom{n_p - i}{\ell - i} (\alpha - 2)^{\ell - i} ,$$

with $\delta = \begin{cases} 0 & \text{if } h-\ell = 0 \\ 1 & \text{otherwise} \end{cases}$, and N_i being the number of vertices in $\mathbf{C}_i[s]$ with $\sum_{i=0}^{n_p} N_i = N - \ell$. Clearly, the quantities $N_i/(N-\ell)$ can be viewed as densities of the non-masters in the i-th compatibility classes. We shall in the sequel assume that these densities are given by $\mid \mathbf{C}_i[s] \mid /(\alpha^n - \mid \Gamma_n[s] \mid)$, i.e. that the non-masters are uniformly distributed among their i-th compatibility classes.

3.3.2. The stationary distributions

In order to compute the stationary distribution (Appendix) we make advantage of the fact that the process considered here is of the form:

$$\begin{aligned} \mathbf{P}_{\ell,\ell+1} &= f(N,\ell,\sigma)\ \Lambda_+ \left(1 + \frac{C_+}{\ell}\right) \quad \text{and} \\ \mathbf{P}_{\ell,\ell-1} &= f(N,\ell,\sigma)\ \Lambda_- \left(1 + \frac{C_-}{N-\ell}\right) \quad , \end{aligned}$$

with $f(N,\ell,\sigma) = (N-\ell)\ell \ / \ \big(N(N-1)\big)$ and

$$\begin{aligned} \Lambda_+ &= \sigma\bar{W}_{\mu,\mu} - \bar{W}_{\nu,\mu} \quad \text{and} \quad C_+ = \frac{(N-1)\,\bar{W}_{\nu,\mu}}{\sigma\bar{W}_{\mu,\mu} - \bar{W}_{\nu,\mu}} \quad , \\ \Lambda_- &= \bar{W}_{\nu,\nu} - \sigma\bar{W}_{\mu,\nu} \quad \text{and} \quad C_- = \frac{(N-1)\,\sigma\bar{W}_{\mu,\mu}}{\bar{W}_{\nu,\nu} - \sigma\bar{W}_{\mu,\nu}} \quad , \end{aligned}$$

respectively[6]. The stationary distribution of the birth-and-death process is obtained from

$$\overline{\Pi}_k = \frac{\pi_k}{\sum_{k=1}^{N} \pi_k} \quad \text{with}$$

$$\begin{aligned} \pi_k = \prod_{\ell=1}^{k} \frac{\mathbf{P}_{\ell-1,\ell}}{\mathbf{P}_{\ell,\ell-1}} &= \frac{B(N, C_-)}{(k + C_+)\, B(1 + C_+, k)\, B(N - (k-1), C_-)} \cdot \\ &\cdot \left(\frac{\sigma\bar{W}_{\mu,\mu} - \bar{W}_{\nu,\mu}}{\bar{W}_{\nu,\nu} - \sigma\bar{W}_{\mu,\nu}}\right)^{k-1} \cdot \frac{N(N-1)\,\bar{W}_{\nu,\mu}}{k\,\big(\sigma(k-1)\bar{W}_{\mu,\nu} + (N-k)\bar{W}_{\nu,\nu}\big)} \end{aligned} \tag{9}$$

[6] $f(N,\ell,\sigma)$ happens to be independent of σ in this particular case. It would, for example, depend on σ if we used the unscaled time axis t instead of $\tilde{t}$.

and $B(z-y,y) = \Gamma(z-y)\Gamma(y) \;/\; \Gamma(z)$ being the Beta and the Gamma function, respectively. The expectation value of the number of master vertices in the long time limit is then obtained by $\mathrm{E}\{\hat{Y}\} = \sum_{k=1}^{N} k\ \overline{\Pi}(k)$.

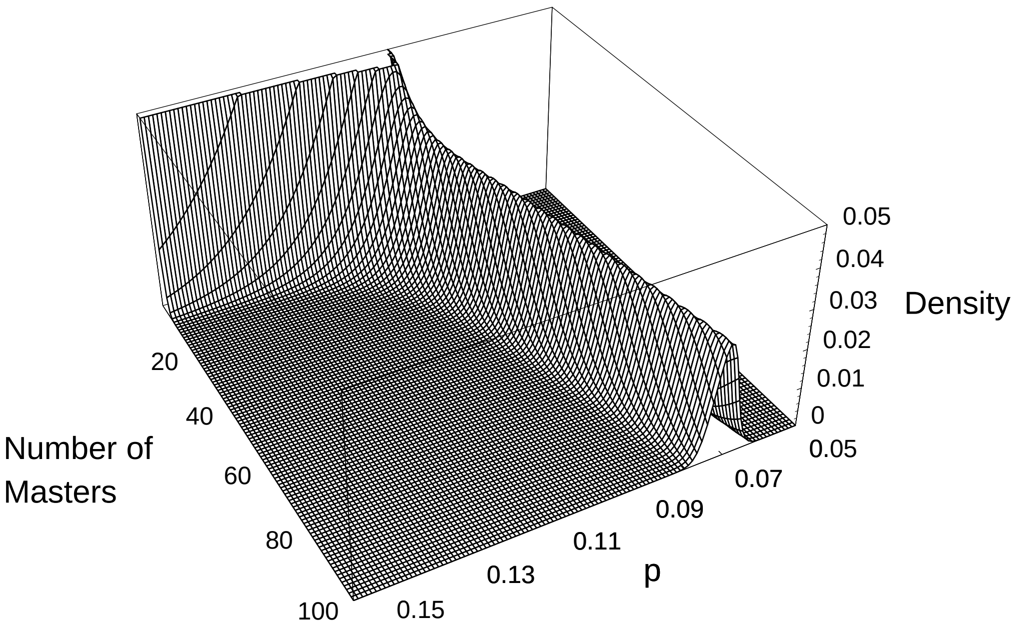
The stationary distribution of the birth-and-death process is evaluated as a function of the mutation rate p. For this goal we approximate the factorials of the Beta function by means of Stirling's formula and plot the stationary distribution as a function of p (figure 2). The maximum of this stationary distribution – although not identical but very close to $\mathrm{E}\{\hat{Y}\}$ – corresponds to the stationary frequency of phenotypes ($\bar{\eta}_m$ in equation 4) shown in figure 1A.

3.3.3. The error threshold for finite populations

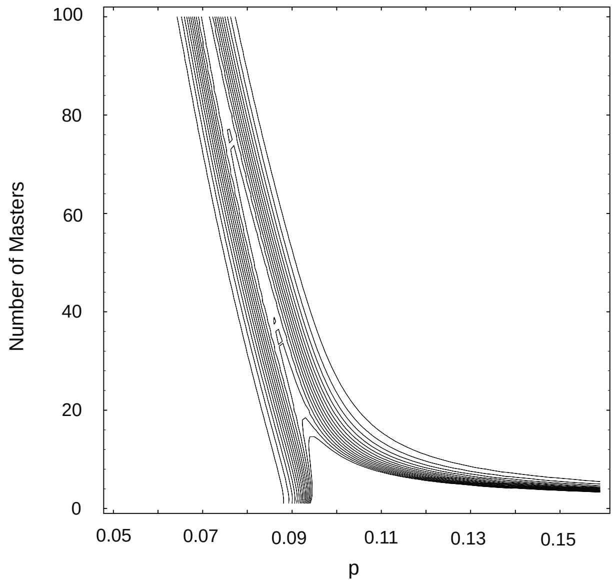
The stationary distribution of the birth-and-death process is applied now to derive an expression for the phenotypic error threshold. Without loosing generality for all realistic situations we assume that the populations size N is small compared to the total number of sequences: $N \ll |\mathcal{Q}_\alpha^n| = \alpha^n$. In addition, it is straightforward to check that the size of a neutral network is usually much smaller than sequence space: $|\Gamma[s]| \;/\; \alpha^n \ll 1$. The number of sequences on the neutral net and in the zero-th compatibility class is: $|\Gamma[s]| + |\mathbf{C}_0[s]| = \alpha^{n_u}\beta^{n_p} \ll \alpha^{n_u+2n_p}$ for sufficiently long sequences and a sufficiently large number of base pairs. Then, the probability $\boldsymbol{\mu}\{\,\hat{Y}=0\,\} = \overline{\Pi}_0$ becomes the maximum of the distribution function in the range of random replication[7] – although $Y=0$ is not an absorbing state.

The stochastic extension of the zero-th order condition for the occurrence of a phenotypic error threshold ($\bar{\eta}_m = 0$) is to ask for the condition that the variance of the stochastic variable $\hat{Y}$ becomes larger than the square of its expectation value in the long time limit. It is meaningful to correct the expectation value by substraction of its minimum value at random replication. Then, we find for the

[7] Random replication means that correct and incorrect digits are incorporated with equal probabilities. It occurs for a replication accuracy in the neighborhood of $q_{rnd} = 1 - 1/\alpha$.



A



B

Figure 2: The stationary distribution of the number of master phenotypes $\bar{Y}_\ell$ as a function of the mutation rate p. The upper part (**A**) shows a 3D-plot of the distribution $\overline{\Pi}_k(p)$ and the lower part (**B**) is a contour plot of **A**. The neutral network $\Gamma[s]$ assumed to be regular and the parameters $\lambda_u{=}\lambda_p{=}0.5$, $\sigma{=}10$ and $N{=}1000$ were used.

phenotypic error threshold of structure s as master phenotype:

$$p^*_{max}(N) \doteq \max \left\{ p \mid \mathrm{Var}\{\hat{Y}_s(p)\} \le \left(\mathrm{E}\{\hat{Y}_s(p)\} - \frac{|\Gamma[s]|}{\alpha^n} \right)^2 \right\} . \tag{10}$$

This criterion for the error threshold appears to be more widely applicable than one previously derived for the genotypic error threshold by means of a birth-and-death model [31]. Attempts to derive an analytical expression for $p^*_{max}(N)$, however, have failed so far. Some numerically computed data for equation (10) are shown in table 3.

3.3.4. The error threshold for infinite populations

In principle, it would be straightforward to extend condition (10) for the error threshold to the limit of large population sizes $p^*_{max} = \lim_{N\to\infty} p^*_{max}(N)$. Since no analytical expression is available, however, a corresponding numerical procedure would bring no deeper insight into the problem. Instead, we derive the corresponding quadratic equation obtained from the kinetic equations (see section 3.2 and [31]), insert the corresponding mean probabilities and find for the frequency of master phenotypes:

$$\bar{\eta}_\mu^2 \; + \; \frac{\sigma \bar{W}_{\mu,\mu} - 1 - \bar{W}_{\nu,\mu}}{\sigma - 1} \, \bar{\eta}_\mu \; - \; \frac{\bar{W}_{\nu,\mu}}{\sigma - 1} \; = \; 0 \; .$$

The frequency of mutant phenotypes is simply obtained from the conservation relation $\eta_\nu = 1 - \eta_\mu$.

Neglecting mutational backflow ($\bar{W}_{\nu,\mu} = 0$) yields the expression

$$\bar{\eta}_\mu(p) \;=\; \frac{\sigma \bar{W}_{\mu,\mu} - 1}{\sigma - 1} \; , \tag{11}$$

which has to be compared with equation (4) derived from the kinetic differential equations. The mean probability $\bar{W}_{\mu,\mu}$ is the analogue to the effective replication accuracy $\tilde{Q}_{mm}$ which we rewrite in order to point at the similarity to equation (3):

$$\tilde{Q}_{mm} \;=\; (1-p)^n \left(\Phi(p)\,\lambda \,+\, 1 \right) \quad \text{with} \quad \Phi(p) \doteq \frac{1}{(1-p)^n} - 1 \; . \tag{3'}$$

This expression is indeed identical with the part in (8) which deals exclusively with unpaired bases. As in section 3.2 we derive the phenotypic error threshold in the infinite population by computing p^*_{max} from the equation

$$\bar{W}_{\mu,\mu}(p^*_{max}) \;=\; \sigma^{-1}\ . \tag{12}$$

In section 4.2 we report numerical results for some concrete examples.

4. Comparison with results of computer simulation

The predictions of the phenomenological and the birth-and-death approach to replication and mutation on neutral networks were checked by means of computer simulations based on Gillespie's algorithm for chemical reactions [18, 19]. The stationary concentrations of master phenotypes and the phenotypic error threshold are studied for a tRNA structure as a realistic example. In case of the stochastic process a model structure of chain length $n = 30$ with 12 unpaired nucleotides and 9 base pairs was chosen for comparison. Stationary concentrations are computed for the master shape and the phenotypes in different compatibilty classes. Values for error thresholds are estimated from the transitions observed in the η_k/p-plots.

4.1. The phenomenological approach

In order to test the phenotypic error threshold derived from the kinetic equations we choose phenylalanyl-tRNA as a special example. The secondary structure of this tRNA (see figure 3) is the phenotype under consideration. The fraction of neutral neighbors of the $\mathrm{tRNA}^{\mathrm{phe}}$ secondary structure is computed by averaging over a sufficiently large number of sequences forming this structure in order to assure convergence. This fraction is resolved for individual positions to test the assumptions of the derivation in section 3.1. The plot in figure 3 shows that a single λ-value for all positions is presumably not a good approximation but two or more different λ-values, one for base pairs and one, two ore more for single stranded nucleotides, may come close to the actual distribution. The chance to obtain a sequence folding into the same structure is apparently smaller for mutations at the nucleotides in base pairs regions than in the single stranded stretches and thus the λ-value is smaller for the nucleotides in base pairs.

In order to put the assignment of λ-values on a quantitative basis we consider several partitionings of the positions in the structure of tRNA$^{\rm phe}$. The results of least square fits are shown in table 1. The large standard deviation of the one λ approach is substantially reduced in the double stranded stretches of the two λ approximation but the standard deviation within the unpaired nucleotides remains still large. The situation is improved by partitioning the unpaired regions into flanking, interior bases, and free dangling ends. The variations become small indeed for the first two classes of single stranded nucleotides. The fraction of neutral neighbors in the free ends still remains high and variable but their influence on the total molecule is small since they occur only in small numbers.

In order to apply the four λ approximation to replication and mutation on the neutral network of tRNA$^{\rm phe}$ we assume independence of the corresponding groups of nucleotides in the sense that the sequence space is factorized into four subspaces. The assumption is the same as the one made in the separation of the spaces of single stranded nucleotides and base pairs in section 2.1. As a consequence of this assumption the contributions of both classes of nucleotides to the effective accuracy factor $\tilde{Q}_{mm}$ are factorizable (equation 8). Accordingly, we apply

$$\tilde{Q}_{mm} \;=\; (1-p)^n \prod_k \Big(\Phi_k(p)\,\lambda_k \,+\, 1\Big) \quad \text{with} \quad \Phi_k(p) \doteq \frac{1}{(1-p)^{n_k}} - 1 \;. \qquad (3")$$

with $k = ('d','f','l','e')$. As follows from table 1 the individual values for tRNA$^{\rm phe}$ are: $(\lambda_d = 0.085, n_d = 40)$, $(\lambda_f = 0.296, n_f = 12)$, $(\lambda_l = 0.583, n_d = 20)$ and $(\lambda_e = 0.801, n_e = 4)$. The stationary fraction of master phenotypes is readily computed as a function of the error rate p from equation 4. The results are shown

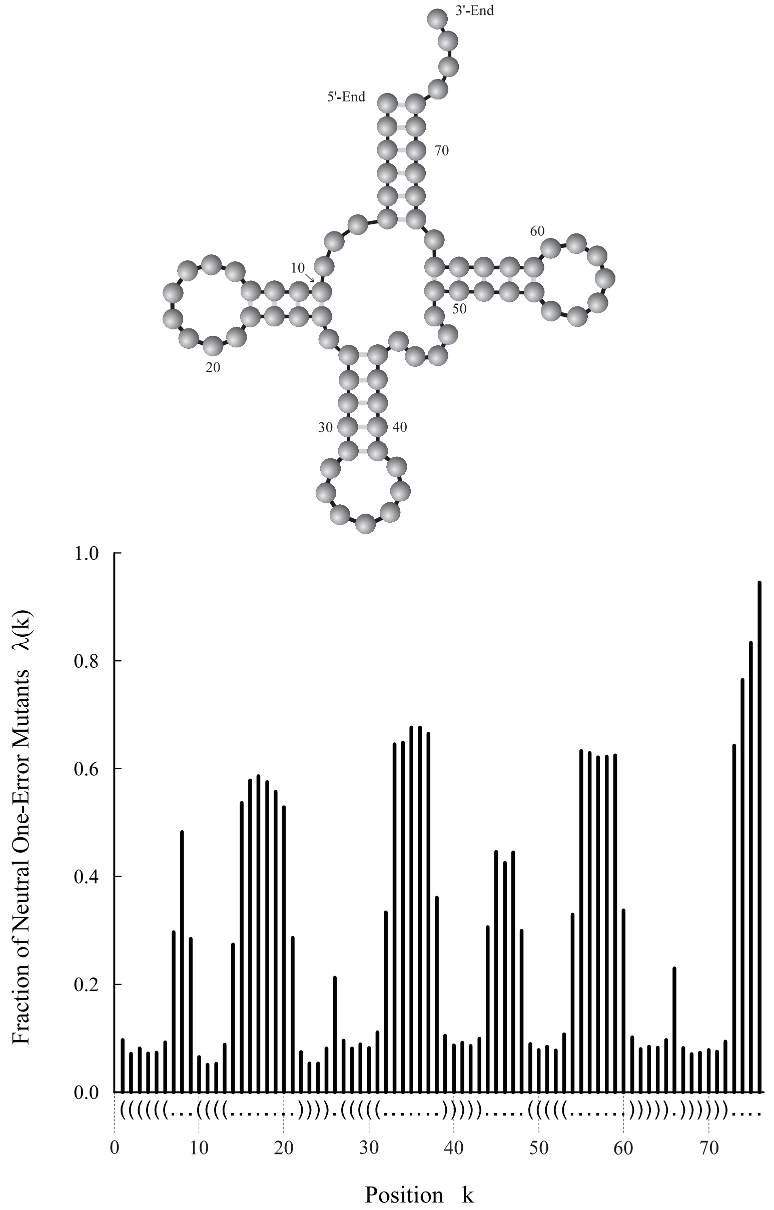


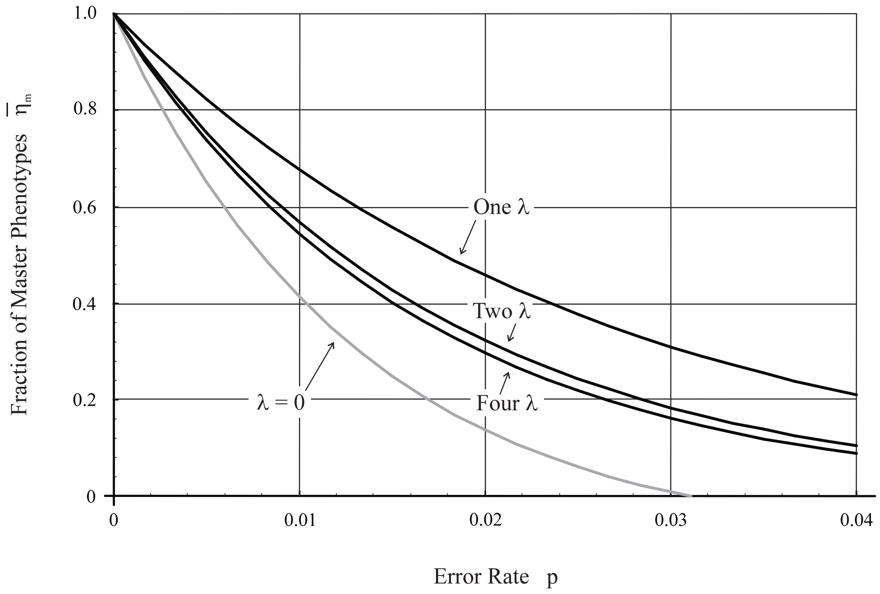
Figure 3: Fraction of neutral variants in the one error neighborhood of the secondary of tRNA$^{\rm phe}$. The secondary structure of tRNA$^{\rm phe}$ is shown in the upper part of the figure. The lower part presents the fraction of neutral neighbors at individual positions k. These fractions are derived by considering all single point mutations of sequences folding into the tRNA$^{\rm phe}$ at minimum free energy conditions. The mean value is obtained by averaging over a sufficiently large number of sequences forming the tRNA$^{\rm phe}$ structure. Below the plot the secondary structure is written in the **symbolic notation**: unpaired nucleotides are represented by dots, "·", and base pairs by parentheses, "(" and ")". The definition of RNA secondary structures (section 2.1) guarantees unique assignment of left and right parentheses to base pairs (see e.g. [11]).

Table 1: Mean fractions of neutral neighbors in tRNA$^{\text{phe}}$ at different levels of resolution. The classification of nucleotides is derived from figure 3.

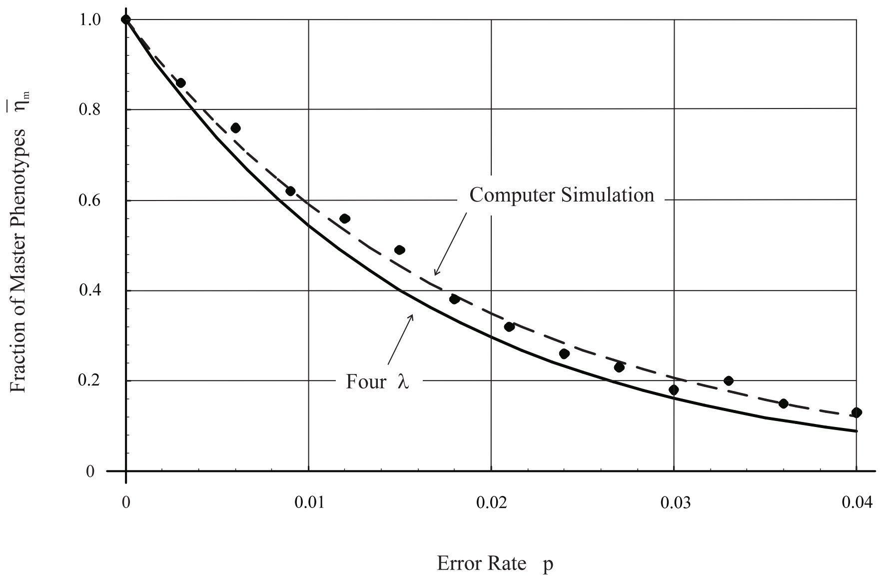
Notation	Positions †	Mean value	Stand. dev.
Single λ approximation			
λ	1–76	0.2871	$\pm\,0.2489$
Two λ approximation			
λ_d	Double helices: (1-6/67-72),(10-13/22-25) (27-31/39-43),(49-53/61-65)	0.0849	$\pm\,0.0148$
λ_s	Single strands: 7-9,14-21,26,32-38, 44-48,54-60,66,73-76	0.5116	$\pm\,0.1841$
Four λ approximation			
λ_d	Double helices: (1-6/67-72),(10-13/22-25) (27-31/39-43),(49-53/61-65)	0.0849	$\pm\,0.0148$
λ_f	Flanks of stacks: 7,9,14,21,26,32,38, 44,48,54,60,66	0.2964	$\pm\,0.0422$
λ_l	Interior of loops: 8,15-20,33-37, 45-47,55-59	0.5830	$\pm\,0.0813$
λ_e	Dangling ends: 73-76	0.8013	$\pm\,0.1270$

† Numbers refer to the conventional positions of nucleotides in tRNA$^{\text{phe}}$. Double helical regions are given in parentheses.

for the different approximations as well as for the "non-neutral" limit ($\lim \lambda \to 0$) in figure 4. The one λ approximation is substantially more different from the two λ curve than the latter from the four λ curve thus indicating convergence of the many λ approximation to the effective replication accuracy $\tilde{Q}$.



A



B

Figure 4: Stationary fraction of tRNA$^{\text{phe}}$ master phenotypes as a function of the error rate p. The upper part (**A**) shows the results of the one-, two- and four λ approximation. The "non-neutral" curve (λ=0, gray) has been added for comparison. In the lower part (**B**) the four λ approximation is compared with the results of computer simulations (individual points and exponential fit, dashed). Values for the superiority σ_m at different error rates p were taken from the results of the simulations (see text).

In order to be able to compare the calculated results with "exact" data, the error propagation problem on the neutral network of tRNA$^{\mathrm{phe}}$ was embedded in a realistic fitness landscape and simulated on the computer. The calculations were performed with a population of about $N \approx 1000$ molecules replicating independently in a flow reactor (for details see [13, 15, 24]). Accordingly, the population size is not precisely constant but fluctuates around the expectation value N with a standard deviation of $\pm\sqrt{N}$. The replication rate is assumed to depend on the shape (but not on the particular sequence): the largest rate constant is assigned to the master shape being the structure of tRNA$^{\mathrm{phe}}$, all other shapes have smaller rate constants according to their distance from the master shape which is measured as the Hamming distance between the two symbolic notations (figure 3). The fitness function is calibrated to yield a superiority of $\sigma_m \approx 10$. Individual simulations were performed for different (constant) values of mutation rates in the range $0 < p \leq 0.04$ and continued until convergence of the fraction of master phenotypes was achieved ($\eta_m(t) \rightarrow \bar{\eta}_m$). Since the stationary population is expanding with incrasing error rate p in sequence space, the superiority of the master shape, σ_m determined after convergence, is not constant. In addition convergence becomes weaker at higher error rates and eventually breaks down because the population cannot cover a section of sequence space which is sufficiently large to support a constant distribution of shapes. As expected from previous studies based on structure autocorrelation functions [11] the mean distance between the master shape and the other shapes in the population becomes larger with increasing width of the distribution. Accordingly, the superiority of the master shape increases at higher mutation rates. A fit to the data reported here yielded:

$$\sigma_m(p) \;=\; 8.369 + 68.37\,p + 4052\,p^2 \quad \text{for} \quad 0 < p \leq 0.04 \;.$$

In the range of p values studied here the superiority σ_m increased by a factor of about two from 8.5 to 17.7. The individual points, $\bar{\eta}_m(p)$, are shown together

Table 2: Phenotypic and genotypic error thresholds, $p_{\max}(N)$, for replication and mutation of tRNA$^{\mathrm{phe}}$ in infinite and finite populations of size N.

Superiority σ	Maximum error rate $p_{\max}(\infty)$	$p_{\max}(1000)$	$p_{\max}(100)$
Phenotypic threshold: four λ approximation			
8.47	0.045	0.042	0.038
10.00	0.049	0.047	0.042
17.67	0.067	0.063	0.057
Genotypic threshold: $\lambda = 0$			
8.47	0.028	0.025	0.021
10.00	0.030	0.027	0.022
17.67	0.037	0.034	0.027

with an exponential fit and the curve computed with the four λ approximation in figure 4 B. The (best) analytical approximation reproduces well the data of the computer simulation: the stationary $\bar{\eta}_m$ values of the simulation are slightly above the values computed from equation 4 in excellent agreement with the neglect of mutational backflow in the analytical approximation.

In table 2 computed error threshold values for different superiorities are compared with the threshold values in absence of selective neutrality. In addition, we show the influence of the population size on the threshold which has been approximated by the simple relation,

$$q_{min}(N) \;=\; q_{min}(\infty)\left(1 + \frac{2\sqrt{\sigma - 1}}{n\sqrt{N}} + \ldots\right) , \tag{13}$$

derived from a stochastic model of replication-mutation kinetics in absence of neutrality [31].

4.2. The birth-and-death model

In order illustrate the results of the stochastic approach and to test its predictions we performed computer simulations for a model molecule of chain length $n = 30$ containing 12 unpaired nucleotides and 9 base pairs. The numerical results are summarized in the two figures 5 and 6 as well as in table 3.

The error threshold at infinite population size is readily computed from equation (12). Its dependence on the degree of neutrality (λ_u, λ_p) is shown in figure 5. The value of the error threshold increases almost linearly with λ. At low λ values an analogous linear dependence is observed in case of the phenomenological threshold (figure 1). The function for the phenotypic threshold, however, levels off and approaches $p_{\max} = 1$ at $\lambda = \sigma_m^{-1}$ whereas $p^*_{\max}$ reaches a maximal value substantially smaller one at $\lambda = 1$. The difference is a consequence of the analytical form of $\Psi(p)$ in equation (8) accounting for contribution of base pairs. Intuitively, the behaviour of the phenomenological threshold appears to be correct since we expect that the phenotype cannot be lost at sufficiently high degree of neutrality. A straightforward explanation for this discrepancy is the apparent difference between the unit distance in the subspace of base pairs $(\mathcal{Q}_\beta^{n_p})$, being the single pair exchange, and the elementary move in mutation, being the change of a single nucleotide.

In order to test the predictions of the birth-and-death model computer simulations were performed. The population size was $n = 1000$ and the shapes were grouped according to the compatibility classes $\mathbf{C}_i[s]$ with the master shape s. The results are shown in figure 6. The curves describing the dependence of the stationary frequencies of the individual compatibility classes on the error rate p are qualitatively identical with the results obtained for error classes in the "non-neutral" case [31]. The four diagrams allow to diagnose the existence and to evaluate the position of the error threshold in straightforward manner. The error threshold becomes less sharp with increasing degree of neutrality as is clearly seen

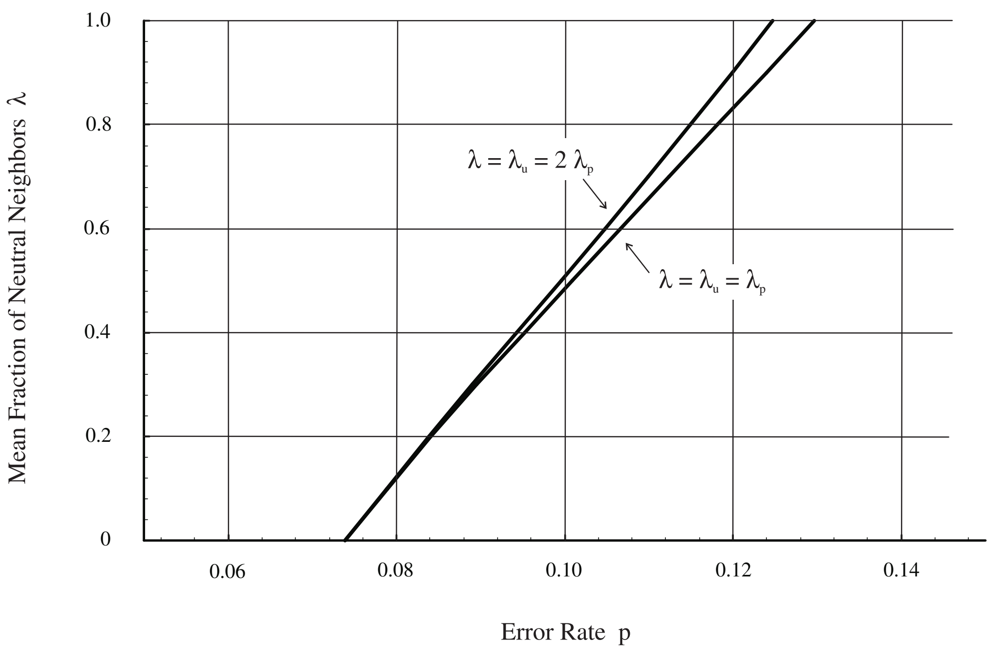


Figure 5: The error threshold at infinite population size. The value of the threshold $p^*_{\max}$ is computed from equation (12) and plotted as a function of the degree of neutrality expressed by the two parameters λ_u and λ_p. The two curves correspond to the relations $\lambda{=}\lambda_u{=}\lambda_p$ and $\lambda{=}\lambda_u{=}2\lambda_p$, respectively. The other parameters in the computation of $\bar{W}_{\mu,\mu}$ according to equation (8) are $n{=}30$, $n_u{=}12$, $n_p{=}9$, and $\sigma{=}10$.

from a comparison of the three cases $\lambda_u = \lambda_p = 0.1$, 0.5, and 0.8. Mixed values, for example $\lambda_u = 0.27$ and $\lambda_p = 0.5$, have thresholds of intermediate sharpness. The four cases studied here show, in combination of the previous observations, that the sharpness of the threshold decreases with increasing $p^*_{\max}$ value. The threshold values in populations of $n = 1000$ molecules were computed from equation (10) and agree well with those derived from the Gillespie simulations. In addition, we estimated the finite population size effects also by applying approximation (13) to the values for infinite populations. These values obtained thereby agree with those calculated from equation (10) within the accuracy limits. Accordingly, accurate estimates of error thresholds in finite populations are possible also by combining equations (12) and (13), and thus allow to circumvent an evaluation

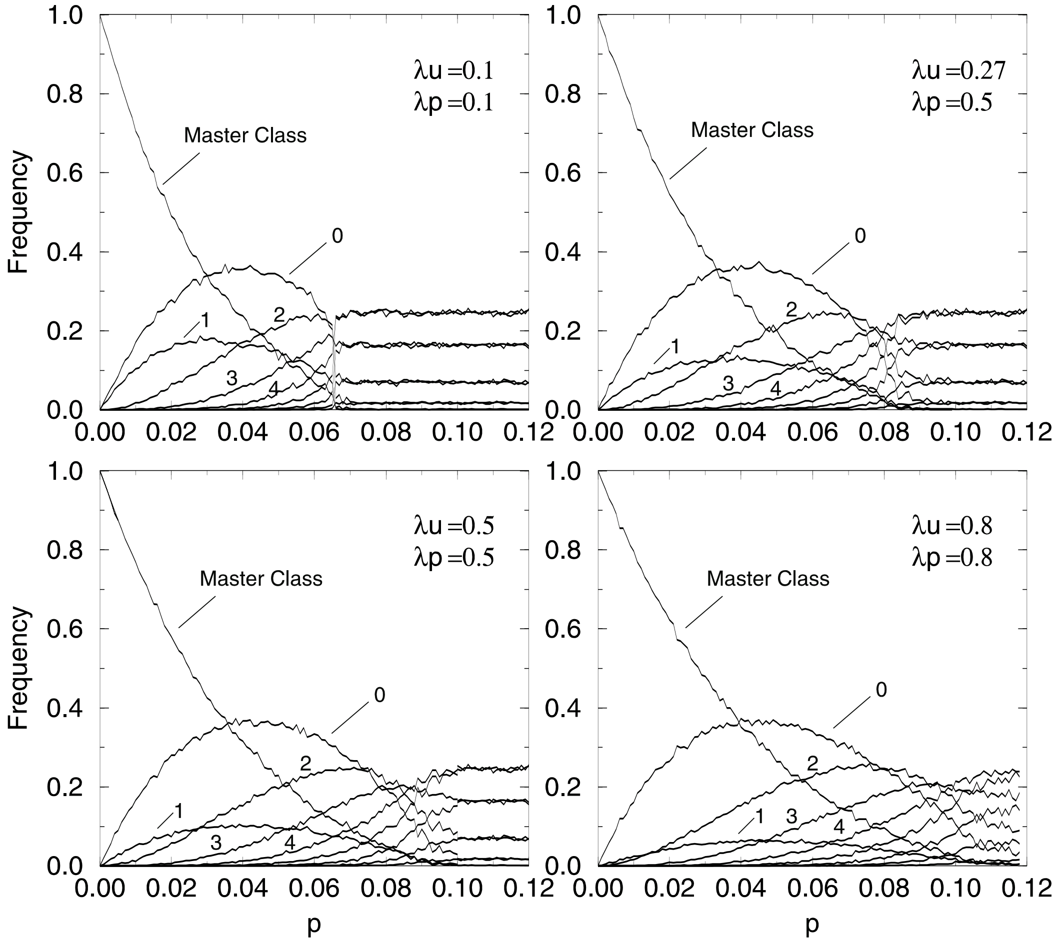


Figure 6: Stationary frequencies of master shapes s and non-master shapes grouped into compatibility classes $\mathbf{C}_i[s]$ for different landscapes (λ_u, λ_p). The stationary frequencies were computed by means of Gillespie's algorithm [18, 19] with the parameters N=1000, n=30, n_u=12, n_p=9, and σ=10.

of equation (10). The population size dependence of stationary concentrations of master phenotypes and error thresholds derived in [31] appears to be of general validity.

Table 3: Numerical values for the phenotypic error threshold in finite and infinite populations. Values obtained from equations (10) and (12) are compared with the results derived from computer simulations using Gillespie's method [18, 19]. A superiority of $\sigma = 10$ was chosen for the master phenotype.

λ_u	λ_p	Birth-and-Death Process		Gillespie Simulation
		Equ.(12) $N = \infty$	Equ.(10) † $N = 1000$	$N = 1000$
0.1	0.1	0.079	0.07 (0.073)	0.065
0.27	0.5	0.081	0.08 (0.075)	0.085
0.5	0.5	0.105	0.10 (0.099)	0.095
0.8	0.8	0.118	0.12 (0.112)	0.110

† Values in parentheses are computed from the error threshold for infinite population size and equation (13).

5. Concluding remarks

Two approaches to handle replication and mutation on neutral networks were developed, analysed and tested: (i) phenomenological differential equations derived from chemical kinetics of replication and mutation on multiple degenerate fitness landscapes and (ii) stochastic processes on neutral networks. Both methods reproduce well the results of computer simulations which were performed as suitable test cases. The phenomenological approach has the advantage of conceptual and methodological simplicity. It can be adapted to describe realistic examples of RNA molecules as it was shown in the $\mathrm{tRNA}^{\mathrm{phe}}$ case study. The stochastic model is based on the random graph method to describe neutral networks of sequence-structure maps. It has the advantage of a rigorous mathematical basis and can be readily extended to other mappings of genotypes onto phenotypes. Both approaches yield identical results in the limit of infinite population size when they are based on the same approximations. The most important approximation, made in the phenomenological approach as well as in the special examples of the stochastic model, is the assumption that non-master shapes are distributed uniformly on neutral networks or over compatibility classes, respectively. In addition, mutational backflow from non-master shapes to the master shape has been neglected in infinite populations. Computer simulations demonstrated that both approximations are well justified (as was found previously in the "non-neutral" case). The two methods presented here provide new tools for modeling neutral evolution and optimization on neutral networks. Both techniques can be applied without modification to cases with high mutations rates and high degrees of neutrality.

Acknowledgments

Computer simulations on population dynamics of $\mathrm{tRNA}^{\mathrm{phe}}$ were performed with the assistence and the software of Dr. Walter Fontana. Constructive criticism and stimulating discussions on replication-mutation dynamics on neutral networks with Drs. Walter Fontana, Ivo L. Hofacker, Martijn Huynen, and Peter F. Stadler are gratefully acknowledged. The work was supported financially by the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung*, Projects No. 10578-MAT and 11065-CHE , by the European Commission, Contract Study PSS*0884, and by the Santa Fe Institute.

Appendix: Mathematical Modeling via Birth Death Processes

A birth-death process is a Markovian process in continuous time [1, 4, 26] whose state space are the integers. Let $P_{\ell,\ell'}(t) = \mu\{X_{t+s} = \ell' \mid X_s = \ell\}$ be independent of s i.e. the process X is *homogeneous.*

For $\lim_{h\to 0}\psi(h) = 0$ for a birth-death process holds:

$$
\begin{aligned}
P_{\ell,\ell+1}(h) &= \mathbf{P}_{\ell,\ell+1}\, h + \psi(h) && \text{for } h \searrow 0,\ \ell \geq 0\\
P_{\ell,\ell-1}(h) &= \mathbf{P}_{\ell,\ell-1}\, h + \psi(h) && \text{for } h \searrow 0,\ \ell \geq 1\\
P_{\ell,\ell}(h) &= 1 - (\mathbf{P}_{\ell,\ell+1} + \mathbf{P}_{\ell,\ell-1})\, h + \psi(h) && \text{for } h \searrow 0,\ \ell \geq 0\\
P_{\ell,\ell'}(0) &= \delta_{\ell,\ell'},\ \mathbf{P}_{0,-1} = 0,\ \mathbf{P}_{\ell,\ell+1},\ \mathbf{P}_{\ell,\ell-1} > 0 && \text{for } \ell > 0.
\end{aligned}
$$

We next state an ergodic theorem [10] that shall imply the existence of a stationary distribution for our birth-death process.

Theorem 1. *Let X_t be a homogeneous Markov process with finitely many states $0,\ldots,N$. If there exists a $0 < t^* < \infty$ such that*

$$P_{i,k}(t^*) > 0 \quad \textit{for} \quad 0 \leq i,k \leq N\,.$$

Then there exists the limits

$$\lim_{t\to\infty} P_{i,k}(t) = P_k \quad \textit{for} \quad 0 \leq i,k \leq N\,.$$

For $i \geq k$ we have

$P_{i,k}(h\,(i-k)) = \sum_\ell P_{i,\ell}(h)\, P_{\ell,k}(h\,(i-(k+1))) \geq P_{i,i-1}(h)\, P_{i-1,k}(h\,(i-(k+1)))$.

Further for any birth-death process we immediately verify by induction on $i-k$ that

$$P_{i,k}(h\,(i-k) \geq \prod_{\ell=0}^{i-k-1} P_{i-\ell,i-1-\ell}(h) > 0\,.$$

Along these lines we further obtain for $k \geq i$

$P_{i,k}(h\,(k-i) \geq \prod_{\ell=0}^{k-i-1} P_{i+\ell,i+1+\ell}(h) > 0$ whence the above theorem applies. The corresponding P_k can be computed by use of the Chapman-Kolmogorov backward equation.

Lemma 1. *Suppose $\Gamma < \mathcal{Q}_\alpha^n$ is a neutral network and $\mathcal{A}$ is a $\star$-alphabet. Suppose that we have a random mapping $v = (x_1, ..., x_n) \mapsto v' = (x'_1, ..., x'_n)$, $v, v' \in \mathrm{v}[\mathcal{Q}_\alpha^n]$ that is defined as follows: We set $x_i = x'_i$ with probability $1-p$ and $x_i \neq x'_i$ with uniform probability p.*

$$\bar{W}_{\mu,\mu} = (1-p)^n \Big(\Phi(p)\,\lambda_u + 1\Big)\Big(\Psi(p)\,\lambda_p + 1\Big) \quad \textit{with},$$

$$\Phi(p) \doteq \frac{1}{(1-p)^{n_u}} - 1 \quad \textit{and} \quad \Psi(p) \doteq \left(\frac{p^2}{(\alpha-1)(1-p)^2} + 1\right)^{n_p} - 1$$

and furthermore

$$\bar{W}_{\nu,\mu} = \sum_{h=1}^{n} \lambda_p\, p^h (1-p)^{n-h} \sum_{\ell=0}^{h} \binom{n_u}{h-\ell} \lambda_u^\delta \sum_{i=0}^{\ell} \frac{N_i}{N-\ell}\, 2^i \binom{n_p - i}{\ell - i} (\alpha-2)^{\ell-i},$$

with $\delta = \begin{cases} 0 & \textit{if } h-\ell = 0 \\ 1 & \textit{otherwise}, \end{cases}$ *and N_i being the number of vertices in $\mathbf{C}_i[s]$.*

Proof. (i) Denoting an error at the unpaired positions with $(-,\ \ )$ and at the paired positions with $(\ \ ,-)$, we can distinguish the following four types: $(+,+), (-,+), (+,-)$ and $(-,-)$. The probability for $(+,+)$ is obvious. For $(-,+)$, we have

$$(1-p)^{n-n_u} \sum_{k=1}^{n_u} \binom{n_u}{k} p^k (1-p)^{n_u - k} \lambda_u = [1-(1-p)^{n_u}]\lambda_u (1-p)^{2n_p}.$$

An error at the paired positions implies for alphabets with unequivocal complementary base pairs that *both* positions have to be changed in order to obtain a compatible pair again. With this information the cases $(+,-)$ and $(-,-)$ are straightforwardly to compute.

For other alphabets it suffices to observe that the probability to obtain a vertex $v \in \mathrm{v}[\Gamma_n[s]]$ by a mutation event at the paired positions is given by

$$\sum_{j=1}^{n_p} \binom{n_p}{j} \left[\frac{p^2}{\alpha-1}\right]^j (1-p)^{2(n_p-j)} = \left[\frac{p^2}{\alpha-1} + (1-p)^2\right]^{n_p} - (1-p)^{n_p}.$$

(ii) For an incompatible configuration with exactly i incompatible positions we now assume an mutation at exactly h positions. There can be $0 \le \ell \le n_p$ errors on the paired positions. In order to obtain a compatible configuration, it is necessary to make an appropriate mutation in each of the i incompatible pairs. There are 2^i different choices to do so with corresponding probability $\frac{1}{\alpha-1}$. The remaining $\ell - i$ errors have to occur pairwise in the other pairs which can be done in $\binom{n_p-i}{\ell-i}(\alpha - 1)^{\ell-i}$ different ways with corresponding probability $(\alpha - 2)\ /\ (\alpha - 1)^2$. This completes the proof of the lemma. ∎

The quantities $N_i/(N-\ell)$, viewed as densities of the non-masters in the i-th compatibility classes, are assumed to be given by $\mid \mathbf{C}_i[s] \mid /(\alpha^n - \mid \Gamma_n[s] \mid)$, i.e. the non-masters are uniformly distributed over their i-th compatibility classes.

Next we compute the stationary distribution of a birth-death process whose birth and death rates can be written as

$$\text{For} \quad 0 \le \ell \le N-1 \quad \mathbf{P}_{\ell,\ell+1} = \left[\frac{f_1(\sigma,\ell,N)}{f_2(\sigma,\ell,N)}\right] \Lambda_1 \, [1 + \frac{C_1}{\ell}]$$

$$\text{For} \quad 1 \le \ell \le N \quad \mathbf{P}_{\ell,\ell-1} = \left[\frac{f_1(\sigma,\ell,N)}{f_2(\sigma,\ell,N)}\right] \Lambda_2 \, [1 + \frac{C_2}{(N-\ell)}] \, .$$

We call birth-death processes with this property P-*processes*.

According to [9, 26] the stationary distribution $\boldsymbol{\mu}_p$ is for $1 \le k \le N$: $\boldsymbol{\mu}_p(k) = \pi_p(k)/\sum_k \pi_p(k)$, where

$$\pi_p(k) = \prod_{\ell=1}^{k} \frac{\mathbf{P}_{\ell-1,\ell}}{\mathbf{P}_{\ell,\ell-1}} \, .$$

In the following proposition we shall make use of the relation $B(z-y,y) = \frac{\Gamma(z-y)\,\Gamma(y)}{\Gamma(z)}$ where $B(x,y)$ is the Beta and $\Gamma(x)$ the Gamma function. The above relation is a classical result and was proved by Dirichlet.

Proposition 1. *Suppose a birth-death process is a* P*-process. Then for* $1 \le k \le N$ *its stationary distribution is determined by*

$$\pi_p(k) = \frac{\mathbf{P}_{0,1}}{\mathbf{P}_{k,k-1}} \, \frac{B(N,C_2)}{(k+C_1)\,B(1+C_1,k)\;B(N-(k-1),C_2)} \, [\frac{\Lambda_1}{\Lambda_2}]^{k-1} \, .$$

Proof. By assumption we can write

$$\begin{aligned}
\mathbf{P}_{\ell,\ell+1} &= \left[\frac{f_1(\sigma,\ell,N)}{f_2(\sigma,\ell,N)}\right] \ \Lambda_1 \ [1+\frac{C_1}{\ell}] \\
\mathbf{P}_{\ell,\ell-1} &= \left[\frac{f_1(\sigma,\ell,N)}{f_2(\sigma,\ell,N)}\right] \ \Lambda_2 \, [1+\frac{C_2}{(N-\ell)}] \, ,
\end{aligned}$$

whence

$$\pi_p(k) = \prod_{\ell=1}^{k-1} \left[\frac{1+\frac{C_1}{\ell}}{1+\frac{C_2}{(N-\ell)}}\right] [\frac{\Lambda_1}{\Lambda_2}]^{k-1} \ \frac{\mathbf{P}_{0,1}}{\mathbf{P}_{k,k-1}} \, .$$

This can be rewritten as

$$\pi_p(k) = \frac{B(N+C_2-(k-1),k-1)}{(k+C_1) \ B(1+C_1,k) \ B(N-(k-1),k-1)} \ [\frac{\Lambda_1}{\Lambda_2}]^{k-1} \ \frac{\mathbf{P}_{0,1}}{\mathbf{P}_{k,k-1}} \, .$$

With $B(x,y)B(x+y,z) = B(y,z)B(y+z,x)$ we conclude

$$B(N+C_2-(k-1),k-1)/B(N-(k-1),k-1) = B(N,C_2)/B(N-(k-1),C_2) \, ,$$

whence the proposition. ∎

References

[1] K. L. Chung. *Markov chains with stationary transition probabilities.* Springer-Verlag, Berlin, 1960.

[2] C. Darwin. The origin of species, 1859. See e.g. Everyman's Library, Vol.811, page 81. J.M.Dent & Sons. London, 1967.

[3] E. Domingo and J. J. Holland. RNA virus mutations and fitness for survival. *Ann.Rev.Microbiol.*, 51:151–178, 1997.

[4] J. L. Doob. *Stochastic Processes.* J.Wiley and Sons, New York, 1953.

[5] M. Eigen. Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 58:465–523, 1971.

[6] M. Eigen. The origin of genetic information. Viruses as models. *Gene*, 135:37–47, 1993.

[7] M. Eigen, J. McCaskill, and P. Schuster. The molecular quasispecies. *Adv. Chem. Phys.*, 75:149 – 263, 1989.

[8] M. Eigen and P. Schuster. The hypercycle. A principle of natural self-organization. Part A: Emergence of the hypercycle. *Naturwissenschaften*, 64:541–565, 1977.

[9] W. Feller. *An Introduction to Probability Theory and its Applications*, volume I and II. John Wiley, New York, London, Sydney, 1966.

[10] M. Fisz. *Wahrscheinlichkeitsrechnung und mathematische Statistik.* VEB Deutscher Verlag der Wiss., Berlin, 1989.

[11] W. Fontana, D. A. M. Konings, P. F. Stadler, and P. Schuster. Statistics of RNA secondary structures. *Biopolymers*, 33:1389–1404, 1993.

[12] W. Fontana, W. Schnabl, and P. Schuster. Physical aspects of evolutionary optimization and adaptation. *Phys. Rev. A*, 40:3301–3321, 1989.

[13] W. Fontana and P. Schuster. A computer model of evolutionary optimization. *Biophys. Chem.*, 26:123–147, 1987.

[14] W. Fontana and P. Schuster. Continuity in evolution. On the nature of transitions. *Science*, 1998. In press.

[15] W. Fontana and P. Schuster. Shaping space: The possible and the attainable in RNA genotype-phenotype mapping. *J. Theor. Biol*, 1998. Submitted. Also published as Santa Fe Institute Preprint No. 97-11-081.

[16] C. V. Forst, C. Reidys, and J. Weber. Evolutionary dynamics and optimization: Neutral networks as model-landscapes for RNA secondary-structure folding-landscapes. In F. Morán, A. Moreno, J. J. Merelo, and P. Chacón, editors, *Advances in Artificial Life*, volume 929 of *Lecture Notes in Artificial Intelligence*, pages 128–147, Berlin, 1995. Springer-Verlag.

[17] C. W. Gardiner. *Handbook of Stochastic Methods, 2nd Ed.* Springer-Verlag, Berlin, 1985.

[18] D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comp. Phys.*, 22:403–434, 1976.

[19] D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, 81:2340–2361, 1977.

[20] U. Göbel, C. V. Forst, and P. Schuster. Structural constraints and neutrality in RNA. In R. Hofestädt, T. Lengauer, M. Löffler, and D. Schomburg, editors, *Proceedings of the German Conference on Bioinformatics 1996*, volume 1278 of *Lecture Notes in Computer Science*, pages 156–165, Berlin, 1997. Springer-Verlag.

[21] W. Grüner, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, and P. Schuster. Analysis of RNA sequence structure maps by exhaustive enumeration. I. Neutral networks. *Mh.Chem.*, 127:355–374, 1996.

[22] W. Grüner, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, and P. Schuster. Analysis of RNA sequence structure maps by exhaustive enumeration. II. Structure of neutral networks and shape space covering. *Mh.Chem.*, 127:375–389, 1996.

[23] I. L. Hofacker, W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Mh. Chem.*, 125:167–188, 1994.

[24] M. A. Huynen, P. F.Stadler, and W. Fontana. Smoothness within ruggedness: The role of neutrality in adaptation. *Proc. Natl. Acad. Sci. USA*, 93:397–401, 1996.

[25] B. L. Jones, R. H. Enns, and S. S. Rangnekar. On the theory of selection of coupled macromolecular systems. *Bull. Math. Biol.*, 38:12–28, 1975.

[26] S. Karlin and H. M. Taylor. *A first Course in Stochastic Processes.* Academic Press, 1975.

[27] M. Kimura. Evolutionary rate at the molecular level. *Nature*, 217:624–626, 1968.

[28] M. Kimura. *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge, UK, 1983.

[29] J. L. King and T. H. Jukes. Non-Darwinian evolution: Random fixation of selectively neutral variants. *Science*, 164:788–798, 1969.

[30] P. A. P. Moran. *The Statistical Processes of Evolutionary Theory.* Clarendon Press, Oxford (UK), 1962.

[31] M. Nowak and P. Schuster. Error thresholds of replication in finite populations. Mutation frequencies and the onset of Muller's ratchet. *J. Theor. Biol.*, 137:375–395, 1989.

[32] C. Reidys, P. F. Stadler, and P. Schuster. Generic properties of combinatory maps - Neutral networks of RNA secondary structures. *Bull. Math. Biol.*, 59:339–397, 1997.

[33] C. M. Reidys. Random induced subgraphs of generalized n-cubes. *Adv. in Appl. Math.*, 19:360–377, 1997.

[34] P. Schuster. Genotypes with phenotypes. adventures in an RNA toy world. *Biophys. Chem.*, 66:75–110, 1997.

[35] P. Schuster. Landscapes and molecular evolution. *Physica D*, 107:351–365, 1997.

[36] P. Schuster. The role of neutral mutations in the evolution of RNA molecules. In S. Suhai, editor, *Theoretical and Computational Methods in Genome Research*, pages 287–302, New York, 1997. Plenum Press.

[37] P. Schuster, W. Fontana, P. F. Stadler, and I. L. Hofacker. From sequences to shapes and back: A case study in RNA secondary structures. *Proc.Roy.Soc.(London)B*, 255:279–284, 1994.

[38] P. Schuster, P. F. Stadler, and A. Renner. RNA Structure and folding. From conventional to new issues in structure predictions. *Curr. Opinion in Struct. Biol.*, 7:229–235, 1997.

[39] S. Spiegelman. An approach to the experimental analysis of precellular evolution. *Quart. Rev. Biophys.*, 4:213–253, 1971.

[40] J. Swetina and P. Schuster. Self-replication with errors - A model for polynucleotide replication. *Biophys.Chem.*, 16:329–345, 1982.

[41] M. Tacker, P. F. Stadler, E. G. Bornberg-Bauer, I. L. Hofacker, and P. Schuster. Algorithm independent properties of RNA secondary structure predictions. *Eur.Biophys.J.*, 25:115–130, 1996.

[42] C. J. Thompson and J. L. McBride. On Eigen's theory of the self-organization of matter and the evolution of biological macromolecules. *Math. Biosc.*, 21:127–142, 1974.

[43] M. S. Waterman. Secondary structures of single stranded nucleic acids. *Adv. Math. (Suppl. Studies)*, 1:167–212, 1978.

[44] T. Wiehe and P. Schuster. Replication-mutation dynamics on different classes of fitness landscapes. Unpublished, 1997.

[45] S. Wright. The roles of mutation, inbreeding, crossbreeding and selection in evolution. In D. F. Jones, editor, *Int. Proceedings of the Sixth International Congress on Genetics*, volume 1, pages 356–366, 1932.

[46] S. Wright. Random drift and the shifting balance theory of mutation. In K. Kojima, editor, *Mathematical Topics in Population Genetics*, pages 1–31, Berlin, 1970. Springer-Verlag.

[47] M. Zuker and P. Stiegler. Optimal computer folding of larger RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9:133–148, 1981.

Table of Contents