Entropic contribution to phenotype fitness

OPEN ACCESS
IOP Publishing
Journal of Physics A: Mathematical and Theoretical
J. Phys. A: Math. Theor. **56** (2023) 345601 (24pp)
https://doi.org/10.1088/1751-8121/ace8d6

Entropic contribution to phenotype fitness

Pablo Catalán[1,2], Juan Antonio García-Martín[2,3], Jacobo Aguirre[2,4], José A Cuesta[1,2,5] and Susanna Manrubia[2,6,*]

[1] Departamento de Matemáticas, Universidad Carlos III de Madrid, Madrid, Spain
[2] Grupo Interdisciplinar de Sistemas Complejos (GISC), Madrid, Spain
[3] Bioinformática para Genómica y Proteómica. Centro Nacional de Biotecnología (CNB-CSIC), Madrid, Spain
[4] Centro de Astrobiología (CAB), CSIC-INTA, Ctra. de Ajalvir km 4, Torrejón de Ardoz, Madrid, Spain
[5] Instituto de Biocomputación y Física de Sistemas Complejos (BIFI), Universidad de Zaragoza, Zaragoza, Spain
[6] Departamento de Biología de Sistemas. Centro Nacional de Biotecnología (CNB-CSIC),Madrid, Spain

E-mail: smanrubia@cnb.csic.es

Received 31 March 2023; revised 11 July 2023
Accepted for publication 19 July 2023
Published 2 August 2023

Abstract

All possible phenotypes are not equally accessible to evolving populations. In fact, only phenotypes of large size, i.e. those resulting from many different genotypes, are found in populations of sequences, presumably because they are easier to discover and maintain. Genotypes that map to these phenotypes usually form mostly connected genotype networks that percolate the space of sequences, thus guaranteeing access to a large set of alternative phenotypes. Within a given environment, where specific phenotypic traits become relevant for adaptation, the replicative ability of a phenotype and its overall fitness (in competition experiments with alternative phenotypes) can be estimated. Two primary questions arise: how do phenotype size, reproductive capability and topology of the genotype network affect the fitness of a phenotype? And, assuming that evolution is only able to access large phenotypes, what is the range of unattainable fitness values? In order to address these questions, we quantify the adaptive advantage of phenotypes of varying size and spectral radius in a two-peak landscape. We derive analytical relationships between the

* Author to whom any correspondence should be addressed.

three variables (size, topology, and replicative ability) which are then tested through analysis of genotype-phenotype maps and simulations of population dynamics on such maps. Finally, we analytically show that the fraction of attainable phenotypes decreases with the length of the genotype, though its absolute number increases. The fact that most phenotypes are not visible to evolution very likely forbids the attainment of the highest peak in the landscape. Nevertheless, our results indicate that the relative fitness loss due to this limited accessibility is largely inconsequential for adaptation.

Keywords: genotype networks, replicator populations, phenotype size, adaptive transitions, RNA folding, toyLIFE, genotype-phenotype maps

(Some figures may appear in colour only in the online journal)

1. Introduction

Our understanding of how genotypes map onto phenotypes, functional pieces and, eventually, whole organisms, has been boosted by studies of simple genotype-to-phenotype (GP) maps (reviews in [1–3]). At odds with a pre-sequencing era view where the mapping between sequence and function was thought to be one-to-one [4], biologically relevant GP maps are many-to-many, with a huge redundancy that has been deeply explored to show, in particular, that a specific phenotype can be achieved from an astronomically large number of genotypes [5].

The set of genotypes that map to a specific phenotype typically form large networks where genotypes are nodes and links represent a mutational move [6–10]. Assigning a unique phenotype to each genotype partitions the space of genotypes into a set of non-overlapping, but linked, phenotypes, and induces a network-of-networks organization in genotype spaces [11]. Knowing the topological features of genotype networks [2, 10] is essential to perform an accurate description of evolutionary dynamics [12]. Phenotype size, defined as the number of genotypes that map onto that phenotype, follows a very skewed distribution, with a small fraction of the largest phenotypes covering most of genotype space; in many cases, the distribution of phenotype sizes is well fit by a lognormal function [13–17]. Both in numerical and empirical studies [13, 17, 18], observed phenotypes are typically large, while most phenotypes are never visited through blind evolutionary searches. A network representation of related genotypes, instead of a phylogenetic tree, can bring out the existence of cycles that reveal parallel or convergent evolution [19]. Also, the degree distribution of genotype networks can be put in direct correspondence with the robustness of a phenotype: the higher the average degree, the lower the effect of mutations, on average. A remarkable feature identified in multiple GP maps is a linear correlation between phenotype robustness (or average degree of the genotype network) and the logarithm of phenotype size [10, 20–23].

Despite its evolutionary relevance, the adaptive effects of phenotype size remain largely unexplored from a formal viewpoint. When we think of 'fitness' of a population, more often than not we recreate the classical fitness landscape that Wright introduced almost a century ago [24]. In this widespread metaphorical representation, devised long before the community became acquainted with the structure and organization of molecular populations, fitness optima corresponded to hilltops in a two-dimensional landscape: adaptation was a parsimonious process that proceeded always uphill and, once mutation-selection equilibrium was attained, populations were forever sitting at the top of the hill. Though this representation cannot, by construction, include the adaptive effects of robustness in phenotype fitness [22] or

environmental variation [25], Wright's fitness landscapes still condition most expectations on the outcome of the evolutionary process [26–28].

The previous criticism notwithstanding, the last two decades have witnessed an increase in the number of works dealing with the effect of phenotype size in adaptation; terms such as entropy, phenotypic redundancy or landscape flatness have been used as synonyms of size. A pioneering work by Schuster and Swetina [29] discussed cases of competition between two phenotypes where the sequence with the highest selective value had a less efficient neighborhood than that with the second largest selective value; they demonstrated that too low a robustness could be fatal at high mutation rates. High mutation rates were also shown to cause the success of a phenotype with lower selective value but higher robustness, in an analytical work that also analyzed the onset of the error catastrophe in this scenario [30]. In a study of spatial gene regulation during development, it was shown that the convergence of finite populations to the maximally fit phenotype was compromised by the multiplicity or entropy of solutions [31]. The survival of the flattest was also considered a surprising effect where a population of replicators would select regions of the landscape of lower fitness but 'flatter', at sufficiently high mutation rates [32]. In a related work where this effect was empirically tested with viroids, the authors stated that fitness should not always be associated with fast replication, and that fitness can indeed be maximized by reducing the impact of mutations on a phenotype [33]. Computational analyses of population dynamics with mutation on the genotype and selection on the phenotype have further clarified the relevance of phenotype size, in phenomena termed the ascent of the abundant [34] or the arrival of the frequent [35, 36].

There is thus broad evidence that evolving populations do tend towards an optimum that is (at least) a combination of replicative ability and phenotype redundancy. In this work, we quantitatively derive the contribution of both terms to the overall fitness under simple conditions. We begin by presenting numerical and theoretical evidence of some important properties of genotype networks, and formally study the case of a population evolving on a network formed by two phenotypes of different size and replicative ability, much in the spirit of [29]. Our aim is to establish the conditions under which the population would transition from one phenotype to the other, and express the transition point as a function of phenotype properties. In order to provide a numerical illustration of the theoretical results, we explore two GP maps of different complexity. First, we use the RNA sequence-to-secondary structure (S3) map, a paradigmatic example [37–39] for which precise numerical and theoretical results regarding the topological nature of its phenotype networks are available. Second, we revisit a pattern-generating version of toyLIFE [36] that we call toyLIFE T2P. toyLIFE is a multilevel GP map that relies on the simple hydrophobic-polar (HP) model for basic interactions [15, 40, 41] and that, despite its complexity, displays qualitative properties analogous to RNA. We close by analyzing and discussing the implications that selection for large phenotypes has in the attainment of phenotypes of sufficiently high replicative ability.

2. Genotype networks

Here, a phenotype is defined as a connected network of genotypes with the same replicative ability. Genotypes are sequences of letters taken from a given alphabet. Two nodes are linked if they differ in one position of their sequences. The number of neighbors of a given node, its degree k_i, is a measure of robustness: the degree is low when point mutations tend to modify the phenotype of sequences one mutation away, while it is high for highly neutral sequences, whose phenotype is typically maintained under mutations. The average degree of a phenotype is defined as $\langle k \rangle = N^{-1} \sum_{i=1}^{N} k_i$, where N is the phenotype size, or the number of nodes in

its genotype network **G**; **G** is the (symmetric) adjacency matrix of a connected (undirected) graph, whose elements are $G_{ij} = 1$ if nodes i and j are connected and $G_{ij} = 0$ otherwise, and whose topology is characterized by a spectral radius γ, the largest eigenvalue of **G**. Finally, we assume a constant environment; otherwise, both the precise set of nodes forming the network and/or the value of the replicative ability could change.

2.1. Evolutionary dynamics of replicator populations

The evolution of a population of replicators on a fitness landscape, assuming discrete generations for simplicity, can be written as

$$\mathbf{n}(t) = \mathbf{M}\mathbf{n}(t-1) = \mathbf{M}^t\mathbf{n}(0) = \sum_{i=1}^{N} \lambda_i^t(\mathbf{n}(0)\cdot\mathbf{u}_i)\mathbf{u}_i \,, \tag{1}$$

where $\mathbf{u}_i$ and λ_i are the eigenvectors and eigenvalues of the evolution matrix **M**, and $\mathbf{n}(t)$ has length N [12, 42]; $\mathbf{n}(0)$ is the initial condition. By definition, the nonnegative matrix **M** is primitive (see below), so the Perron–Frobenius theorem ensures that, over time, the system evolves towards an asymptotic state characterized by the unique first (in decreasing order of eigenvalues) eigenvector $\mathbf{u}_1$. In biological terms, this state corresponds to the mutation-selection equilibrium. The components of $\mathbf{u}_1$ are all strictly positive and proportional to the asymptotic fraction of the total population at each node, while its associated eigenvalue λ_1 represents the asymptotic growth rate of the population.

In a population of replicators that mutate with probability $0 < \mu < 1$ per genotype and replication cycle, matrix **M** can be decomposed as

$$\mathbf{M} = (1-\mu)\mathbf{R} + \frac{\mu}{S}\mathbf{G}\mathbf{R}, \tag{2}$$

where **R** is the diagonal matrix $R_{ij} = r_i\,\delta_{ij}$, r_i being the replicative ability of node (genotype) i. For a fixed phenotype, we will consider in this contribution that $r_i \equiv r$ for all i, where r can be interpreted as the average number of copies of a given sequence in the next generation (time step). In other words, we are considering a single phenotypic trait for all sequences, so that the population is monomorphic. S stands for the maximum number of neighbors of a genotype. When replicators are sequences of length L whose elements are taken from an alphabet of $A \geqslant 2$ letters, the size of the genotype space is $m = A^L$, and $S = L(A-1)$.

Matrices such as **M** in equation (2) are guaranteed to be primitive if the network **G** is connected and the diagonal of **R** is strictly positive. Both conditions are fulfilled, by definition and because $r_i > 0$ represent replicative values.

Matrices **M** and **G** share eigenvectors (because $\mathbf{R} = r\mathbf{I}$), and their respective eigenvalues λ and γ are related through

$$\lambda = r\left[(1-\mu) + \frac{\gamma\mu}{(A-1)L}\right] . \tag{3}$$

2.2. Competition between phenotypes

Consider two phenotypes α and β, each represented by a different network, with parameters $N_{\alpha/\beta}$, $r_{\alpha/\beta}$ and $\gamma_{\alpha/\beta}$. Further assume that the two phenotypes are mutually accessible through single mutations (excluding deletions and insertions) from one or a few nodes in their networks (see figure 1). The matrix **M** describing the evolution of a population of replicators in the two-peak landscape formed by the two phenotypes has a diagonal term for replication plus a topological contribution: two blocks along the diagonal, each corresponding to one of

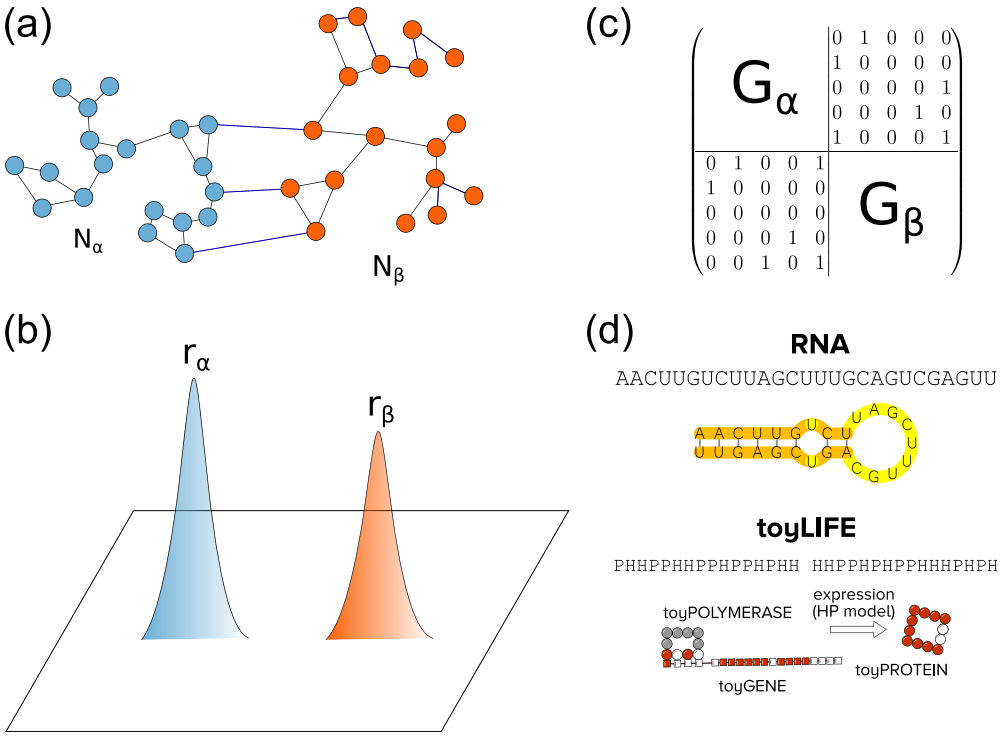

Figure 1. Schematic representation of the scenario considered in this work. A phenotype is characterized by a neutral network of genotypes, all of which share the same replicative ability r. (a) We consider a situation with two phenotypes, α and β, with networks of sizes N_α and N_β and known topology, on (b) a two-peak landscape representing replicative abilities. (c) The joint adjacency matrix of the $\alpha+\beta$ system is formed by two diagonal blocks, each containing the adjacency matrix $\mathbf{G}_\alpha$, $\mathbf{G}_\beta$ of each phenotype (with spectral radii γ_α and γ_β) plus a number of off-diagonal terms that represent mutational connections between the two phenotypes. (d) The phenotypes of two GP maps will be computationally studied, RNA S3 and toyLIFE T2P. See main text for further details.

the phenotypes, and one or a few non-zero elements off the diagonal blocks representing the connections between nodes in different phenotypes just one mutation away. Specifically,

$$\mathbf{M}_{\alpha+\beta} = (1-\mu)\mathbf{R}_{\alpha+\beta} + \frac{\mu}{S}\mathbf{G}_{\alpha+\beta}\mathbf{R}_{\alpha+\beta}\,, \tag{4}$$

with $\mathbf{G}_{\alpha+\beta}$ as depicted in figure 1(c); $\mathbf{M}_{\alpha+\beta}$ has dimension $(N_\alpha+N_\beta)^2$, $\lambda_{\alpha+\beta}$ is its largest eigenvalue, and $\mathbf{G}_{\alpha+\beta}$ and $\mathbf{M}_{\alpha+\beta}$ only share eigenvectors when $r_\alpha = r_\beta$.

The question of which of the two phenotypes would be preferred by the population, and the point where most of the population would transition from one phenotype to the other was addressed in a similar scenario by Schuster and Swetina [29]. As a first approximation, let us assume, following those authors, a situation where the two blocks are weakly coupled, say that a single link exists between two typical nodes of phenotype α and phenotype β and $N_\alpha \geqslant N_\beta \gg 1$. In this case, the block (phenotype) with the largest eigenvalue will be the preferred choice and the other one will be asymptotically empty (as $N_\alpha > N_\beta \to \infty$).

This transition is remarkably sudden even for finite and relatively small values of the size of phenotypes [43–45]. That is, if a population mostly occupies phenotype α, it will transition to phenotype β when $\lambda_\beta > \lambda_\alpha$: the eigenvalue λ can be interpreted as the fitness of a phenotype. Note that, in this interpretation, fitness λ results from a non-trivial combination of the topological properties in figure 1(a) and the replication rate, as represented in figure 1(b). Indeed, recalling equation (3), we obtain a relationship between the replication rate and the spectral radii of the two phenotypes, stating that the transition $\alpha \to \beta$ will occur if

$$\frac{r_\beta}{r_\alpha} > \frac{1-\mu+\gamma_\alpha\mu(A-1)^{-1}L^{-1}}{1-\mu+\gamma_\beta\mu(A-1)^{-1}L^{-1}} \simeq 1 + \frac{\mu}{(A-1)L}(\gamma_\alpha - \gamma_\beta), \tag{5}$$

the last approximation holding for $\mu \ll 1$.

If both phenotypes are mutually accessible, the population at equilibrium will be distributed across both phenotypes. The inequality above will accurately quantify the transition point as long as the connectivity between the two phenotypes is weak. As the number of connections between the phenotypes increases, the division of the adjacency matrix into two distinct blocks progressively blurs, the description in terms of the independent eigenvalues worsens and the population transition between phenotypes becomes smoother [43]. Whenever $r_\alpha \neq r_\beta$, the mutation rate μ affects the position of the transition point and the relative fraction of population at each phenotype. We will not explore systematically the effects of μ in this work, and will assume μ is sufficiently small so that equation (5) is a good approximation to the transition point. In that case, the effect of changing the mutation rate is akin to modifying sequence length L or alphabet size A. For example, higher μ values will result in more weight for the effect of the spectral values on the transition; thus, the transition from α to β will occur for lower values of r_β, if $\gamma_\beta > \gamma_\alpha$. In the numerical examples to be discussed later $\mu = 0.1$, while A and L take different values depending on the GP map studied; theoretical approximations and numerical results are in good agreement, as it will be shown.

2.3. Bounds to the spectral radius

The spectral radius is a measure of a network's topology. Its value admits various bounds as a function, in particular, of the average $\langle k \rangle$, maximum $k_{\max}$ and minimum $k_{\min}$ degree in the network [46],

$$k_{\min} \leqslant \langle k \rangle \leqslant \gamma \leqslant k_{\max}\,, \tag{6}$$

equalities holding for homogeneous networks, where all nodes have the same degree.

2.3.1. Spectral radius and the mean degree of a graph. The bound $\langle k \rangle \leqslant \gamma$ is a known result that can be easily proven. The spectral radius of a real symmetric matrix $\mathbf{A}$ is defined as

$$\rho(\mathbf{A}) = \max_{\|\mathbf{x}\|=1} \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}. \tag{7}$$

Thus, if $\mathbf{A}$ satisfies the conditions of Perron–Frobenius's theorem (i.e. if the underlying graph is connected and not multipartite) and $\mathbf{u}$ is the (unique) eigenvector associated to $\rho(\mathbf{A})$ (the largest eigenvalue), then

$$\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} < \mathbf{u}^{\mathrm{T}}\mathbf{A}\mathbf{u} = \rho(\mathbf{A}), \qquad \mathbf{x} \neq \mathbf{u}. \tag{8}$$

So, if $\mathbf{G}$ is the adjacency matrix of a connected graph and γ is its spectral radius, and if we take for $\mathbf{x}$ a uniform vector with components $N^{-1/2}$, then

$$\mathbf{x}^{\mathrm{T}}\mathbf{G}\mathbf{x} = \langle k \rangle < \gamma \tag{9}$$

as long as $\mathbf{x} \neq \mathbf{u}$ (i.e. the graph is not regular).

We can improve this lower bound for γ by repeating the argument with $\mathbf{G}^2$ rather that $\mathbf{G}$. As $\mathbf{x}^{\mathrm{T}}\mathbf{G}^2\mathbf{x} = \|\mathbf{G}\mathbf{x}\|^2$ and $(\mathbf{G}\mathbf{x})_i = k_i N^{-1/2}$, then we obtain $\langle k^2 \rangle < \gamma^2$ for any nonregular graph. If σ^2 denotes the variance of the degree distribution, then

$$\langle k^2 \rangle = \langle k \rangle^2 + \sigma^2 = \langle k \rangle^2 \left(1 + \frac{\sigma^2}{\langle k \rangle^2} \right),$$

and the lower bound becomes

$$\gamma > \langle k \rangle \sqrt{1 + \frac{\sigma^2}{\langle k \rangle^2}} \geqslant \langle k \rangle + \frac{\sigma^2}{2\langle k \rangle^2} \tag{10}$$

(the last inequality follows from the inequality $\sqrt{1+x} \geqslant 1 + (x/2)$, valid for all $x \geqslant 0$). Inequalities (9) and (10) yield important relationships between the average degree of a graph and its spectral radius, which determines the asymptotic state of a population of replicators, as described above. For homogeneous graphs, where $k_i = k = \langle k \rangle$ for all i, $\langle k \rangle = \gamma$ holds. The question is, how far from homogeneous are genotype networks? Are the previous relationships relevant to predict the evolutionary behavior of a population on these networks?

2.3.2. Average degree and network size. GP maps have been broadly used to generate genotype networks and to characterize the topological properties the map confers to sequence spaces [1–3, 47]. Some of the quantities derived, most often numerically, seem to be quasi-universal, in the sense that they are repeatedly found in a variety of GP maps. Such is the relationship between the average degree of a genotype network and its size, which is largely independent of the specific definition of phenotype: $\langle k \rangle \sim \log N$.

This relationship can be heuristically calculated taking as example the case of the RNA sequence-to-secondary structure (S3) map [10], though the results are more general. First, we recall that an excellent estimation of the number N of genotypes folding into a given RNA secondary structure can be obtained by calculating the so-called versatility of each position along the sequence. The versatility v_j of site j, $j = 1, \ldots, L$, is defined as the number of mutations (out of the total size A of the alphabet) that site j accepts, averaged over all sequences in the network [16, 48, 49], from which the size of the phenotype is estimated as

$$N = \prod_{j=1}^{L} v_j. \tag{11}$$

Numerical comparison between this estimation and the exhaustive enumeration of genotypes in a phenotype yields an excellent agreement [16, 49]. Asymptotically, the size of RNA secondary structures admits a two-versatility approximation [14, 16] that distinguishes just two different structural elements, paired and unpaired nucleotides, each class admitting on average a number v_p and v_u of neutral mutations, respectively (see also [50, 51]). In a previous contribution, it was shown that $\langle k \rangle \propto \log N$ using this approximation.

Nevertheless, an argument justifying the dependence of the average degree on $\log N$ can be obtained directly from (11) if we interpret this expression as the product of L random variables: v_i, the versatility of site i for a given phenotype, can be considered an instance of a random variable with a definite distribution, that we assume to have finite variance. By taking logarithms, $\log N$ can be regarded, in the limit $L \to \infty$, as a normal random variable

$$\log N \approx L\nu_{\log} + L^{1/2}\sigma_{\log}\xi, \qquad \xi \sim \mathcal{N}(0,1) \tag{12}$$

where $\nu_{\log} \equiv \langle \log v \rangle$, $\sigma_{\log} \equiv \langle (\log v - \nu_{\log})^2 \rangle$. On the other hand,

$$\langle k \rangle = \sum_{j=1}^{L} (v_j - 1), \tag{13}$$

is (asymptotically in L) another normal random variable

$$\langle k \rangle \approx L(\nu - 1) + L^{1/2} \sigma \xi', \qquad \xi' \sim \mathcal{N}(0,1), \tag{14}$$

where $\nu \equiv \langle v \rangle$, $\sigma \equiv \langle (v - \nu)^2 \rangle$. Although referring to an average as a random variable may sound strange, one should bear in mind that 'average' here is understood over the nodes of a given phenotype. Thus, $\langle k \rangle$ takes different values for different phenotypes, and it is in this sense that should be considered a random variable.

Now, from (12), we can eliminate $L^{1/2}$ as

$$L^{1/2} \approx \left(\frac{\log N}{\nu_{\log}} \right)^{1/2} - \frac{\sigma_{\log} \xi}{2\nu_{\log}} + O\big((\log N)^{-1/2} \big), \tag{15}$$

which implies

$$L \approx \frac{\log N}{\nu_{\log}} - \frac{\sigma_{\log} \xi}{\nu_{\log}^{3/2}} (\log N)^{1/2} + O(1). \tag{16}$$

Substituting these two expressions in (14), we obtain

$$\langle k \rangle \approx \frac{\nu - 1}{\nu_{\log}} \log N + \left(\frac{\sigma}{\nu_{\log}^{1/2}} \xi' + \frac{(\nu - 1)\sigma_{\log}}{\nu_{\log}^{3/2}} (-\xi) \right) (\log N)^{1/2} + O(1), \tag{17}$$

which can be written in the form

$$\langle k \rangle \approx c \log N + \kappa (\log N)^{1/2} \eta, \qquad \eta \sim \mathcal{N}(0,1), \tag{18}$$

where the coefficients c and κ depend on statistical properties of the distribution of versatilities in the GP map, but not on the particulars of a specific phenotype (that information is hidden in the random variable η).

We can estimate the coefficients in the expression (18) using maximum likelihood to fit a normal distribution to the empirical data of a set of P phenotypes. The resulting formulas for them are

$$c = \frac{\sum_{i=1}^{P} \langle k_i \rangle}{\sum_{i=1}^{P} \log N_i}, \qquad \kappa^2 = \sum_{i=1}^{P} \frac{\big(\langle k_i \rangle - c \log N_i \big)^2}{P \log N_i}. \tag{19}$$

Figure 2 shows the fit of equation (18) to RNA S3, $L = 16$, and toyLIFE T2P, with very good results. These also extend previous numerical analysis of RNA sequences of length $L = 12$, which showed that equation (18) yields, to first order in $\log N$, an excellent fit to numerical results (see figure 3(b) in [10]).

In view of the success of the versatility model (11) in describing the size distribution of different GP maps [16], equation (18) turns out to be more general than the above derivation, with the RNA model in mind, might suggest. Independent analyses have shown that the proportionality between average degree and phenotype size is not limited to RNA S3, as it has been numerically obtained in simple models of protein folding [20], in a model for protein quaternary structure [21] and in toyLIFE [22]. Interestingly, it also describes well some empirical observations, as the relationship observed in a genotype network reconstructed from short haplotypes in the human chromosome 22 [23]. Altogether, these results strongly suggest

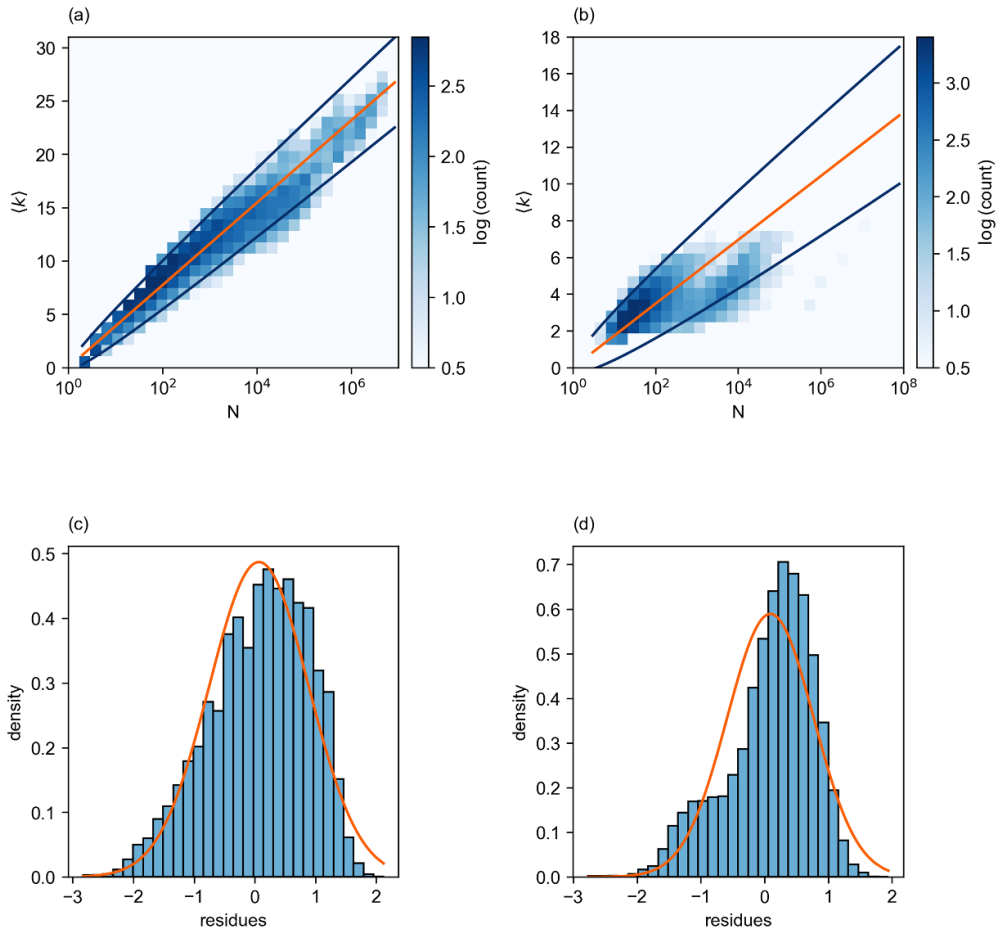

Figure 2. Relationship between $\langle k\rangle$ and $\log N$, in (a) RNA S3 $L=16$ and (b) toyLIFE T2P. The orange line represents the average of $\langle k\rangle$ as given by equation (18), $\nu_{\langle k\rangle} = c\log N$, while the dark blue lines represent the 95% confidence interval, $\nu_{\langle k\rangle} \pm 1.96\sigma_{\langle k\rangle}$, with $\sigma_{\langle k\rangle} = \kappa(\log N)^{1/2}$. For RNA (a), $c = 3.863, \kappa = 0.822$, (a) while, for toyLIFE, $c = 1.7375, \kappa = 0.684$ (b). Lower panels represent the distribution of residues of the least-squares fit $(\log N)^{-1/2}\langle k\rangle = c(\log N)^{1/2}$ for (c) RNA and (d) toyLIFE. Failure to fit to a Gaussian distribution might arise from the small length of genotypes or, especially in toyLIFE, perhaps be a constitutive property of the model.

that $\langle k\rangle \propto \log N$ may be a quasi-universal property of biologically realistic GP maps and fundamentally related to the distribution of phenotype sizes, as the derivation of both results in the framework of the versatility model strongly suggests.

3. Genotype network topology in numerical GP maps

The theory derived in the previous section relates, on the one hand, the fitness λ of a phenotype with its replicative ability and its spectral radius, equation (3), and, on the other hand, the average degree $\langle k\rangle$ and the log size of a phenotype, equation (18). Both expressions are further related through the inequality $\langle k\rangle \leqslant \gamma$.

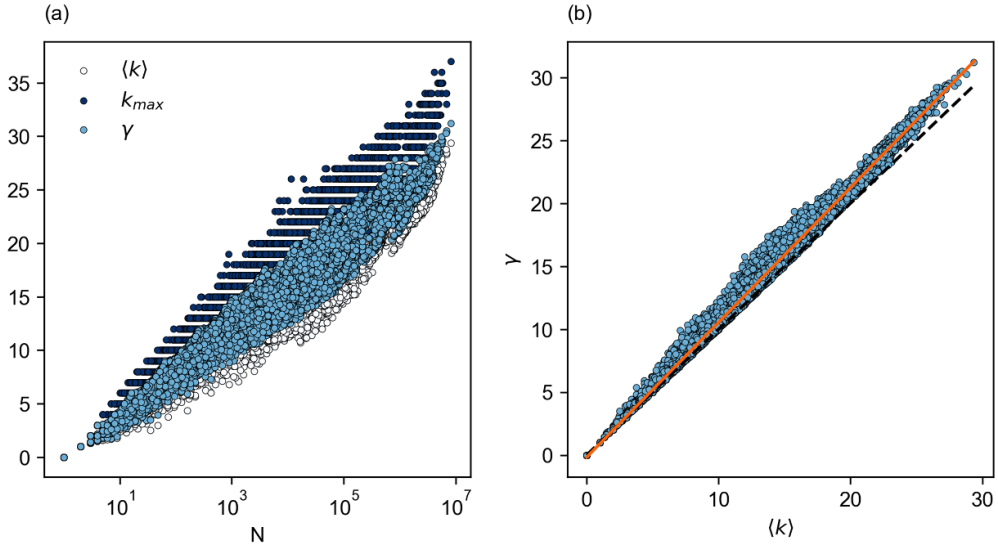

Figure 3. Topological quantities characterizing RNA S3 networks for $L = 16$. (a) We represent the maximum and average network degree, as well as the network spectral radius, for connected components (CC) obtained through the exhaustive enumeration of the sequence space, as a function of CC size. (b) Connected component spectral radius γ as a function of the corresponding average degree $\langle k \rangle$. The orange line represents a linear fit between the two measures: $\gamma = 1.07\langle k \rangle - 0.17$, with $R^2 = 0.99$. The dashed black line represents the line $\gamma = \langle k \rangle$, as a visual aid to confirm that $\gamma \geqslant \langle k \rangle$.

Numerical simulations in this section are devoted to explore the topological properties of two representative GP maps of different complexity, RNA S3 and toyLIFE T2P. Our eventual aim is to check how close the average degree $\langle k \rangle$ is to the spectral radius γ of a genotype network; should the approximation $\langle k \rangle \simeq \gamma$ be feasible, we could derive an approximate relationship between the fitness of a phenotype and its size.

3.1. RNA

We have exhaustively folded the space of RNA sequences of lengths $L = 14, 15$, and 16, mapped each sequence to its minimum-free-energy secondary structure, and separated each phenotype into connected components (CCs) [10, 16]. The results obtained are comparable for the three genotype lengths above (with 3311, 8792 and 23 091 CCs, respectively) and consistent with those obtained for $L = 12$ [10]. Each CC is a connected graph for which we have calculated the maximum degree $k_{\max}$, the average degree $\langle k \rangle$ of its nodes, and the spectral radius γ. Note that the maximum degree $k_{\max}$ corresponds to the node with the largest number of neutral neighbors in each CC and has to fulfill $k_{\max} \leqslant (A-1)L$. The three quantities are jointly represented in figure 3(a) for $L = 16$. In all cases, since all CC fulfill the conditions of the Perron–Frobenius theorem, the inequalities of equation (6) hold.

Figure 3(b) depicts the calculated spectral radius as a function of the average degree. As it can be seen, both quantities are not only proportional, but also remain close for all values of $\langle k \rangle$ represented. Still, numerical data show a persistent dispersion due to specific (non-independent) phenotypic features that affect the degree distribution, such as size or the

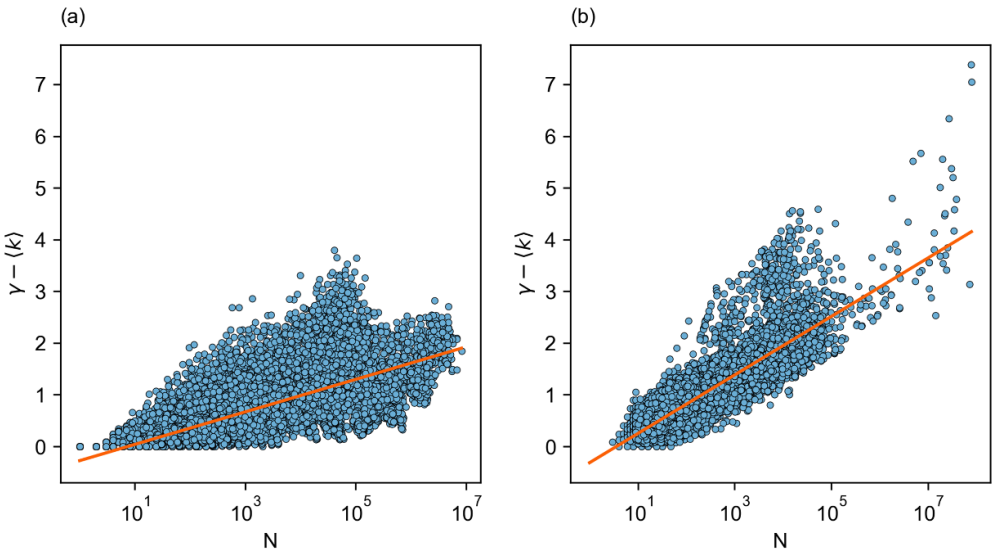

Figure 4. Difference between the spectral radius γ and the average degree $\langle k\rangle$ as a function of network size N for (a) RNA S3 with $L=16$ and (b) toyLIFE T2P. The lowest value $\gamma-\langle k\rangle=0$ corresponds to homogeneous networks, where all nodes have the same degree. In both cases, the orange line represents a linear fit: $\gamma-\langle k\rangle=0.31\log_{10}N-0.28$ for RNA S3 ($R^2=0.56$) and $\gamma-\langle k\rangle=0.57\log_{10}N-0.32$ for toyLIFE T2P ($R^2=0.76$).

total number of paired nucleotides and their distribution within the considered RNA structure. Equation (10) made this dispersion explicit, giving a bound to the difference between both quantities, $\gamma-\langle k\rangle>\sigma^2/(2\langle k\rangle)$. Figure 4(a) illustrates the difference in the case of RNA, showing as well an increase with phenotype size N.

In the limit $L\to\infty$, the distribution of structural elements in RNA secondary structures converges to a Gaussian distribution [14, 52, 53]. This fact does not eliminate the heterogeneity of the network for a fixed (typical) phenotype, but implies that the dispersion σ is similar for different (typical) phenotypes. In this limit, since c becomes independent of the phenotype, there is an additional approximation to the inequality in equation (5) that can be performed. Substituting equation (18), we obtain

$$\frac{r_\beta}{r_\alpha}>1+\frac{c\mu}{(A-1)L}\log\left(\frac{N_\alpha}{N_\beta}\right)+O(\sqrt{\log N})\,. \tag{20}$$

3.2. *toyLIFE*

toyLIFE is a multilevel map from binary genomes, $A=2$, to Boolean gene regulatory networks (GRNs) [40, 41]. toyLIFE sequences code for genes that are translated into 2D compact proteins following the rules of a HP model for protein folding [54]. Each gene is a binary sequence of length 20 with a promoter region of 4 positions plus 16 positions coding for a protein, yielding 4×4 compact lattice proteins (figure 5(a)). These proteins interact to form dimers and, jointly, they alter the expression of genes, thus yielding Boolean GRNs. In Boolean GRNs, genes are represented as being either ON or OFF, and we model them in discrete time, where

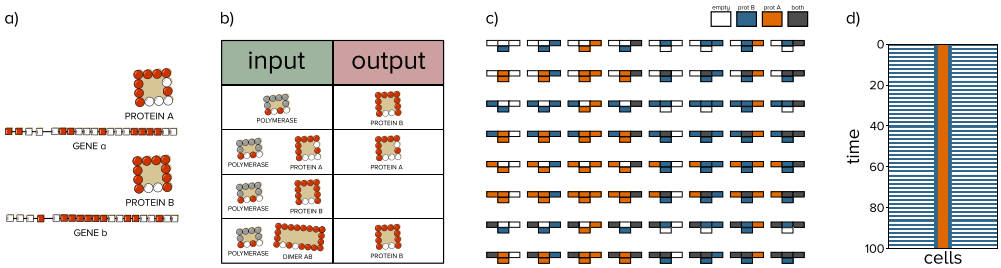

Figure 5. An illustration of the multilevel genotype-phenotype map of toyLIFE with two genes. (a) toyLIFE genotypes are binary strings with a length of $20n$, where n represents the number of genes in the genome (in this case, $n = 2$). Each gene consists of a promoter region represented by the first four letters and a coding region represented by the remaining 16 letters. When the coding region is expressed, it produces a protein that folds into a compact shape on a 4×4 lattice. (b) By following toyLIFE's interaction rules [15], the corresponding gene regulatory network (GRN) is obtained. The GRN is represented here by its truth table. (c) Each GRN, allowing for the diffusion of some proteins to nearest neighbors, determines a unique cellular automaton. At time t, based on the state of a cell and its neighbors, toyLIFE's rules determine the state of the cell at time $t + 1$. The cell can be empty (white), express protein A (orange), express protein B (blue), or express both proteins (grey). (d) When certain initial conditions are met, such as the continuous expression of protein A in the middle cell of the tissue, the cellular automata generate spatio-temporal patterns of gene expression. In this example, the cellular automaton described in (c) produces an alternating pattern where the tissue expresses protein B and then remains inactive, while in the center of the tissue, three cells continuously express protein A. Figure adapted with permission from [36].

the expression of a cell at time $t + 1$ depends on its own expression and that of its neighboring cells at time t (figure 5(b)). This modeling approach transforms GRNs into cellular automata (figure 5(c)). By connecting multiple cells in a one-dimensional tissue and allowing the propagation of proteins between neighboring cells, spatio-temporal patterns similar to those observed in real organisms can be obtained (figure 5(d)). Therefore, toyLIFE serves as a multilevel map from binary genomes to Boolean GRNs to cellular automata to spatio-temporal patterns, enabling the study of molecular evolution at different phenotypic levels. In this work, we define phenotype to be each unique spatiotemporal pattern generated by toyLIFE genomes, such as the one shown in figure 5(d)). We restrict ourselves to the case with two genes (T2P). Hence, $L = 40$ and $k_{\max} \leqslant 40$.

We have selected toyLIFE as a limit example of a complex, yet tractable, GP map that, in its two-gene version, might be severely affected by finite size effects. Still, the qualitative behavior of topological quantities of toyLIFE T2P phenotype networks is equivalent to that described for RNA S3, though the multilevel nature of toyLIFE T2P yields a larger dispersion in the relationships measured. This can be quantitatively observed in figure 6(a), which represents the values of $k_{\max}$, $\langle k \rangle$ and γ for the 12 051 440 CC analyzed in toyLIFE T2P. Note that these are not all CC in T2P, only those obtained from phenotypes with $N < 10^8$ for computational efficiency). Figure 6(b) depicts the relationship between the average degree and the spectral radius. This latter figure indicates that only relatively small CC are homogeneous, since there are no instances of large CC fulfilling the equality $\gamma = \langle k \rangle$. As expected, the observed dispersion is smaller in RNA S3, where all CC remain closer to the diagonal, as shown in figure 3(b).

Finally, we represent in figure 7(a) and b several degree distributions in the two GP maps studied to illustrate their variation with the specific phenotype for finite L, even in phenotypes

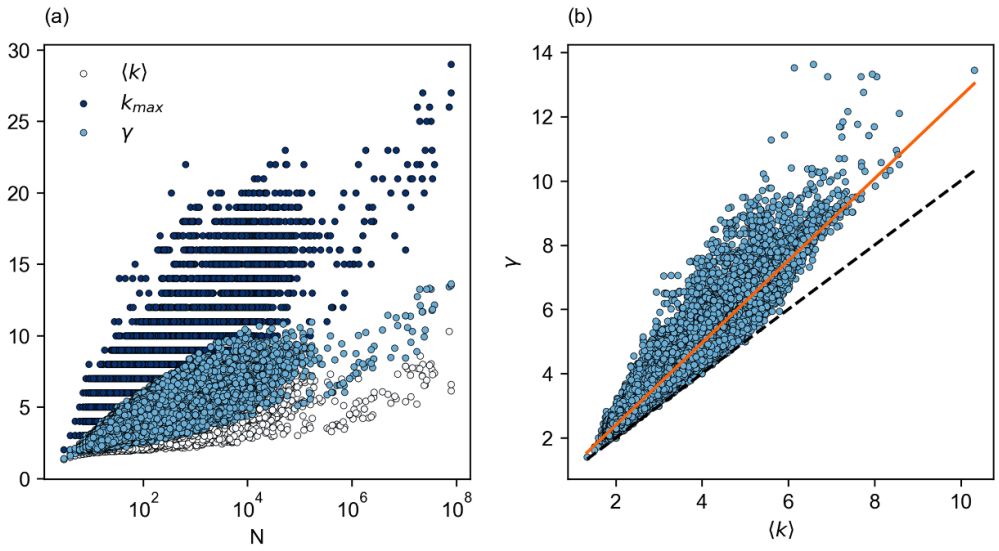

Figure 6. Topological quantities characterizing toyLIFE two-gene networks with a pattern-generating phenotype (T2P). (a) We represent the maximum and average network degree, as well as network spectral radius as a function of CC size, for CCs obtained through the exhaustive enumeration of the sequence space. (b) Connected component spectral radius γ as a function of the corresponding average degree $\langle k \rangle$. The orange line represents a linear fit between the two measures: $\gamma = 1.28\langle k \rangle - 0.16$, with $R^2 = 0.87$. The dashed black line represents the line $\gamma = \langle k \rangle$, as a visual aid to confirm that $\gamma > \langle k \rangle$.

of comparable size. The obtained distributions are relatively peaked around a well-defined average, so the correction obtained by including σ is small. This is further illustrated in figure 7(c) and d, where the bound given by equation (10) is depicted. There is no noticeable improvement with respect to the results reported in figure 4 when the dispersion of the degree distribution is included in the bound.

4. Numerical examples of phenotypic transitions

Estimating the eigenvalue λ of genotype networks in fitness landscapes is difficult for at least two reasons. First, an exhaustive enumeration of all genotypes in a given phenotype is out of reach even for relatively short sequences; second, even if the replicative ability of genotypes is known, the calculation of the spectral radius of large networks is a costly computational procedure. The use of the log size of the phenotype as a proxy for the average degree $\langle k \rangle$, first, and then for γ seems feasible in the light of the numerical results in the previous section. Under these consecutive assumptions, equation (20) estimates the transition point between two phenotypes given their replicative abilities and their sizes. Though this estimation will be necessarily worse than that obtained through λ, it informs on the transition point in many situations where the full degree distribution of a genotype network is not available, or when the coefficients involved in a relationship such as equation (18) are unknown. Therefore, the

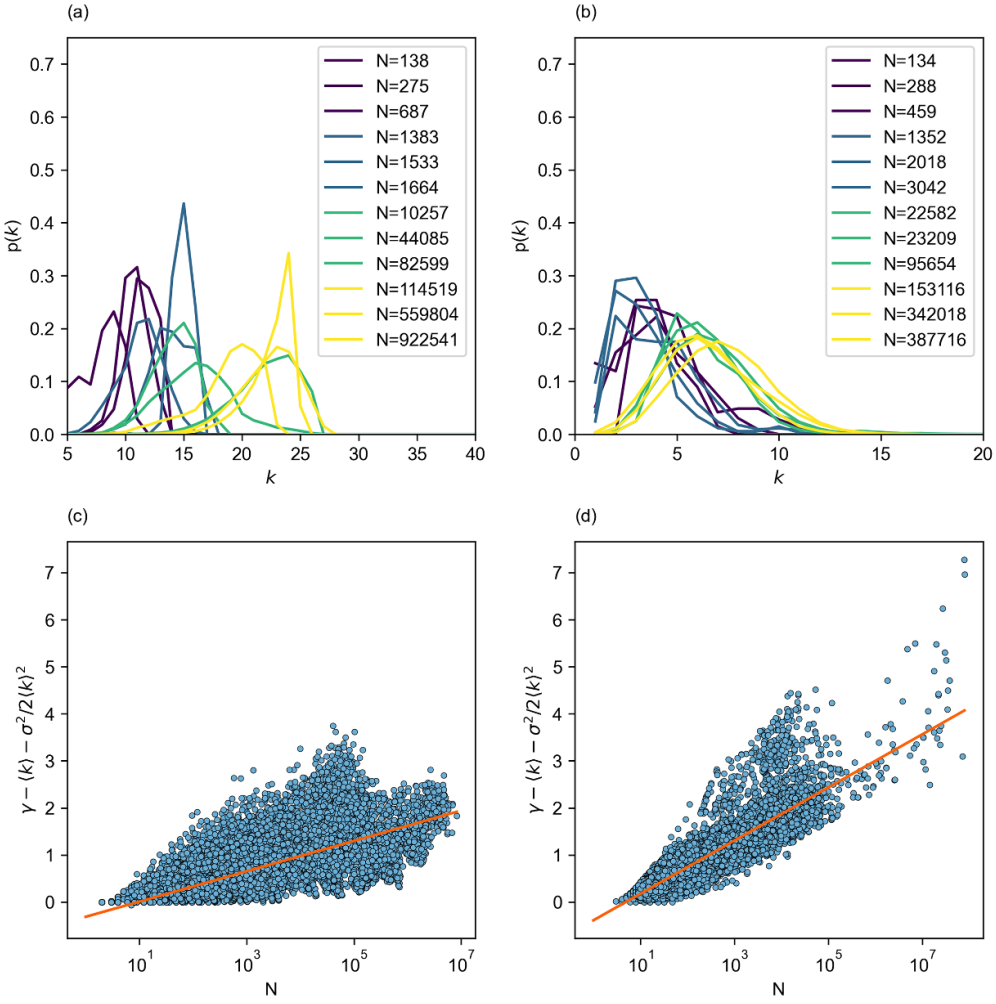

Figure 7. Degree distributions for RNA S3 and toyLIFE T2P and their effects on bounds to the spectral radius. (a) Degree distribution for RNA S3 with $L = 16$; (b) degree distribution for toyLIFE T2P; difference $(\gamma - \langle k \rangle - \sigma^2/2\langle k \rangle^2$ for (c) RNA S3 with $L = 16$ and (d) toyLIFE T2P. In (c) and (d), the orange line represents a linear fit: $\gamma - \langle k \rangle - \sigma^2/2\langle k \rangle^2 = 0.32 \log_{10} N - 0.32$ for RNA S3 ($R^2 = 0.56$) and $\gamma - \langle k \rangle - \sigma^2/2\langle k \rangle^2 = 0.56 \log_{10} N - 0.39$ for toyLIFE T2P ($R^2 = 0.79$).

advantage of using $\log N$ instead of γ comes from the existence of various low-cost computational methods that allow accurate [18] and approximate [16, 49] estimations of phenotype size.

We have explored transitions between phenotypes in RNA S3 and toyLIFE T2P in a two-peak landscape to characterize the transition and, chiefly, to quantify how the transition point depends on the topological characteristics of the phenotype. A schematic of the scenario studied is represented in figure 1. Our exhaustive enumeration of the two GP maps described (RNA

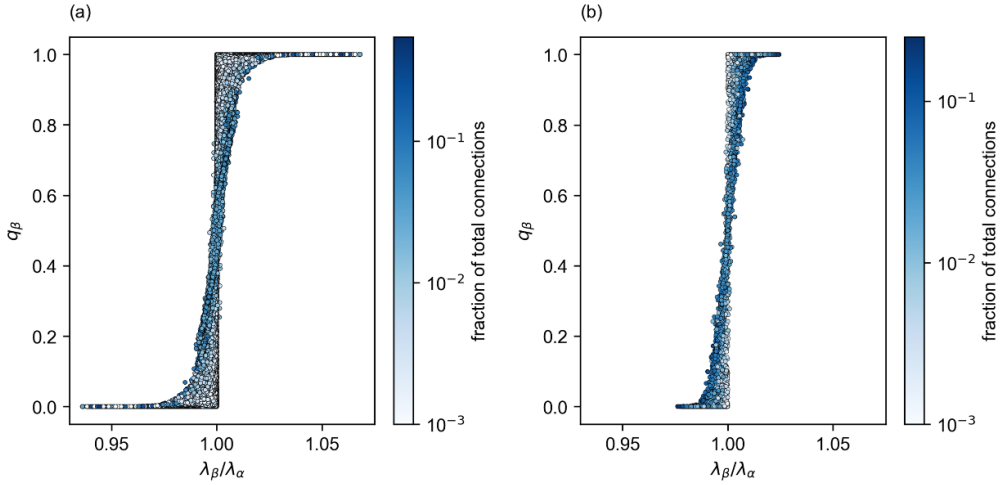

Figure 8. Accuracy of the prediction $\lambda_\alpha = \lambda_\beta$ as transition point. (a) RNA. Each point represents a system of two phenotypes of different size, each characterized through its fitness λ. The vertical axis shows q_β, the fraction of replicators in phenotype β. All 206 821 pairs of the 3311 CCs of RNA S3 with $L = 14$ that are mutually accessible through point mutations were used for this plot. (b) As previous panel, for toyLIFE. The color scale represents the fraction of links that connect every pair of phenotypes. Most pairs of phenotypes follow the prediction, with a dispersion of about 1% around $\lambda_\alpha = \lambda_\beta$. The parameters of the simulations are replicative ability $r_\alpha = r_\beta$, mutation probability $\mu = 0.1$, alphabet length $A = 4$ and $L = 14$ for RNA; $r_\alpha = r_\beta$, $\mu = 0.1$, $A = 2$ and $L = 40$ for toyLIFE. λ_α and λ_β were calculated with equation (3), where γ is the spectral radius of each network. The latter were numerically computed for each network.

S3 and toyLIFE T2P) allows to characterize all phenotype networks and the links between phenotypes, that is, nodes that belong to each of the phenotypes, and the complete set of neighboring phenotypes one mutation away—with at least one connecting pair, but few to many in general. The transition matrix $\mathbf{M}_{\alpha+\beta} = (1-\mu)\mathbf{R}_{\alpha+\beta} + (\mu/S)\mathbf{G}_{\alpha+\beta}\mathbf{R}_{\alpha+\beta}$, with $\mathbf{G}_{\alpha+\beta}$ as depicted in figure 1(c); $\mathbf{M}_{\alpha+\beta}$ has dimension $(N_\alpha + N_\beta)^2$ and $\lambda_{\alpha+\beta}$ is its largest eigenvalue, with associated eigenvector $\mathbf{u} = (u_i)$.

The first approximation we made was that matrix $\mathbf{M}_{\alpha+\beta}$ would have a block-like structure, with few, off-diagonal, non-zero elements, following [29]. In an updated representation [43] the two-phenotype system has been described as two connected networks 'competing' for 'resources' (in the present case, resources correspond to replicators). It has been shown [43] that the eigenvector centrality of connector nodes, in our case the set of genotypes that are one mutation away but belong to different phenotypes, determines how sharp is the transition at $\lambda_\alpha = \lambda_\beta$. The larger the number of connector nodes and their eigenvector centrality, the less accurate becomes the prediction based on the two-block separation.

Let us define $q_\alpha = U^{-1}\sum_{i=1}^{N_\alpha} u_i$ as the fraction of the population of replicators that occupies nodes in phenotype α at equilibrium. Also, there is a fraction $q_\beta = U^{-1}\sum_{i=N_\alpha+1}^{N_{\alpha+\beta}} u_i = 1 - q_\alpha$ of the population of replicators occupying nodes in phenotype β, with $U = \sum_{i=1}^{N_{\alpha+\beta}} u_i$. Figure 8 shows, for many different pairs (α, β) of phenotypes, the fraction q_β at equilibrium. Most of the population occupies nodes in phenotype β when $\lambda_\beta > \lambda_\alpha$, and vice versa. The color scale represents the fraction of links that actually connect both phenotypes in

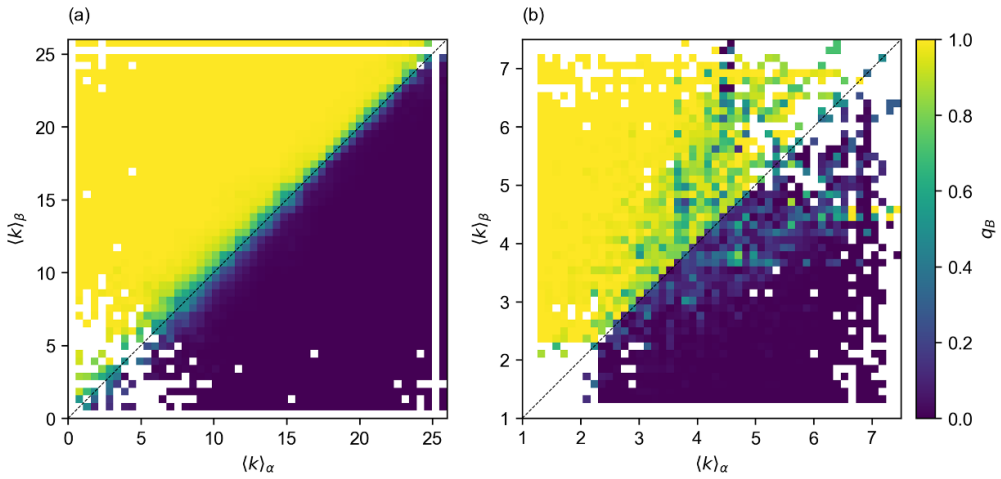

Figure 9. Accuracy of the prediction $\langle k\rangle_\alpha = \langle k\rangle_\beta$ as transition point (assuming $r_\alpha = r_\beta$). Results are represented in the plane $(\langle k\rangle_\alpha, \langle k\rangle_\beta)$. (a) RNA S3, $L = 14$. Each point stands for an average over all pairs of phenotypes with the corresponding average degrees. The color scale indicates the fraction q_β of the total population occupying nodes of phenotype β, averaged over pairs. (b) toyLIFE T2P. As in the previous plot.

relation to its maximum possible number, showing that this fraction is in the majority of cases small. As expected, the transition around $\lambda_\alpha = \lambda_\beta$ is sharp for pairs of phenotypes weakly connected, while large fractions of connector links between phenotypes smoothen the transition [43].

4.1. Transition as a function of average degree

This is a second approximation where we assume that the average degree is a good approximation of the spectral radius, $\langle k\rangle \simeq \gamma$. In previous sections, we have seen that this is not always the case since, though degree distributions are peaked around well-defined average values, genotype networks are heterogeneous in degree, and their heterogeneity does not vanish with increasing phenotype size. What is more important, the average degree depends on each specific phenotype, as we have seen explicitly with RNA, and numerically with the two examples we have explored. An advantage of using the average degree to predict the transition point, however, is that $\langle k\rangle$ can be obtained through suitable sampling of nodes in a genotype network, and does not require an exhaustive knowledge of the network—which is needed to calculate γ or λ.

Figure 9 represents the fraction q_β averaged over all pairs of phenotypes with degree $(\langle k\rangle_\alpha, \langle k\rangle_\beta)$. White points stand for non-existing pairs; statistics are better for average values of the degree, between 5 and 20 for RNA S3, $L = 14$ and 3 and 5 for toyLIFE T2P. The prediction worsens close to the diagonal $\langle k\rangle_\alpha = \langle k\rangle_\beta$, though it is quite good for RNA and slightly worse for toyLIFE. We do not observe any improvement in the predicted transition point with larger average degree. The prediction is very good when the difference between the average degree of the two phenotypes is about 2–3 or larger. Despite all the caveats, the average degree yields a reasonable, probabilistic estimate, of the position of the transition between two phenotypes.

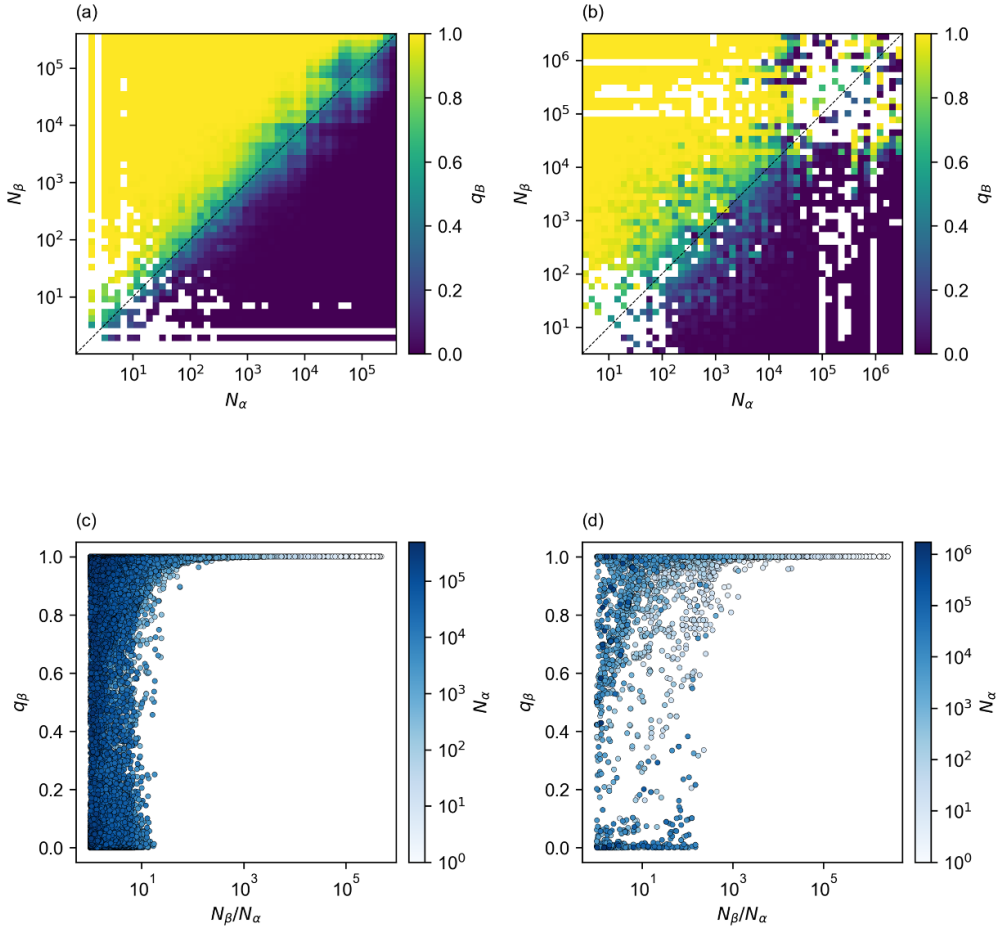

Figure 10. Accuracy of the prediction $\log N_\alpha = \log N_\beta$ as transition point (assuming $r_\alpha = r_\beta$ and a phenotype independent coefficient c). (a), (b) Results are represented in the plane $(\log N_\alpha, \log N_\beta)$ for (a) RNA S3, $L = 14$ and (b) toyLIFE T2P. Each point stands for an average over all pairs of phenotypes with the corresponding sizes. The color scale indicates the fraction q_β of the total population occupying nodes of phenotype β, averaged over pairs. (c), (d) Fraction of population in phenotype β, q_β as a function of the relative size between the two phenotypes, N_β/N_α. The color scale indicates the size of the smallest phenotype in the compared pair. (c) RNA S3, $L = 14$; (d) toyLIFE T2P.

4.2. Transition as a function of phenotype size

The third and last approximation we check here is the substitution of the average degree of a network by the logarithm of the phenotype size, times a phenotype-dependent multiplicative term, $\langle k \rangle \simeq c \log N$ to predict the transition point. As discussed above, there are reasons to assume that c becomes asymptotically independent of the phenotype for typical phenotypes, at least in RNA. We cannot discard that some GP maps may behave otherwise though, as of yet, we do not have examples contradicting that assumption. Figures 10(a) and (b) represents the fraction q_β averaged over all pairs of phenotypes with sizes $(\log N_\alpha, \log N_\beta)$, while figures 10(c) and (d) represents individual pairs. This prediction could be improved had we

included the value of the coefficient c for different phenotypes; however, we have chosen to represent the case where c is assumed to be phenotype-independent as a limit case with the minimum number of quantities to estimate: just phenotype size.

The use of N as a proxy to estimate transitions between phenotypes has a practical and a conceptual implication, as we have anticipated in previous sections. On the practical side, phenotype size can be easily estimated with a variety of methods of different accuracy available in the literature [16, 18, 49]; on the conceptual side, the relationship between the transition point and phenotype size provides a quantitative measure of the importance of phenotype redundancy, a measure of entropy, in phenotype fitness. Though this prediction is not as good as the one obtained if the whole genotype network is known, it is reasonable attending to the computational effort needed to estimate phenotype size. Further, it allows a first estimation of the relative adaptive value of replicative ability versus phenotype size. For RNA S3 and $L = 14$, an order of magnitude difference in phenotype size means that the smaller phenotype is essentially empty; whether this difference remains constant or increases with L remains to be explored. For toyLIFE T2P, the difference required is slightly larger, about 1.5 orders of magnitude. For pairs of phenotypes more similar in size, however, the larger one typically attracts over 50% of the population—note that light-blue points are rare in any case.

5. On finding a sufficiently fit phenotype

Evolution is severely conditioned by the size of phenotypes, as all studies with synthetic and empirical GP maps have demonstrated. In the sections above, we have derived quantitative relationships between the fitness λ and two important features: replicative ability and phenotype size. The dependence with phenotype size N provides a first explanation of why large phenotypes are the only ones seen by natural selection: when a population has to choose between two phenotypes of comparable replicative ability, the larger one will be preferred. The question arises: how much replicative ability is lost as a consequence of phenotype size?

Let us address this question in a simple situation where the distribution of phenotypes sizes is known and fitness values for each phenotype are drawn at random from a distribution of well-defined average, regardless the size of the phenotype. Note that quantitative results of previous sections are not used explicitly here. Instead, we will consider another quasi-universal property of GP maps that explains the huge span in phenotype sizes: the distribution of phenotype sizes is in most cases well fit by a lognormal function [13, 16, 22, 48]. This feature of phenotype sizes entails that there are orders of magnitude difference between abundant, typical, and rare phenotypes, even for relatively short sequences: an astronomically large number of phenotypes invisible to evolution. The preference for large phenotypes can be dramatically illustrated with the case of natural, non-coding RNA sequences [13]. For example, most abundant phenotypes in sequences of length $L = 126$ have sizes between 10^{20} and 10^{40}. However, phenotypes selected in natural systems (meaning here RNA S3 sequences available at the fRNAdb [55]) have sizes not smaller than 10^{36}, reaching 10^{46} in many cases [13]. For other functional, non-coding RNAs subjected to strong selective pressures on the secondary structure, such as viroids [56, 57], phenotype sizes can reach 10^{90} for $L \simeq 399$ [58]. The number of compatible genotypes rapidly becomes hyperastronomically large for any realistic functional phenotype [5].

5.1. Visible values of phenotype size

Although the following discussion applies to any GP map with a log-normal distribution of phenotype size, for the sake of illustration—and because it is the best documented example

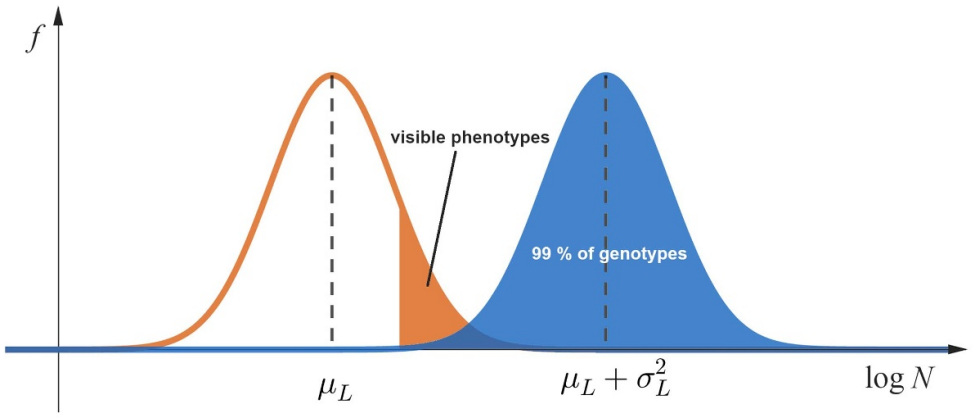

Figure 11. Fraction of phenotypes f with a log-size $\log N$ (orange), as well as the fraction of genotypes in phenotypes of log-size $\log N$ (blue), for sequences of length L. The former are normally distributed with mean μ_L and standard deviation σ_L; the latter are distributed as $Nf(\log N)$—hence as a normal of mean $\mu_L + \sigma_L^2$ and standard deviation σ_L [16]. The 'visible' phenotypes are those in which nearly all genotypes (here 99 %) are concentrated.

in the literature—, we will focus on the case of the RNA S3 map. For this map we know [16] that the fraction of phenotypes with a given $\log N$ is a normal distribution with mean μ_L and standard deviation σ_L, given by

$$\mu_L = \mu_1 L + O(1), \qquad \sigma_L = \sigma_1 L^{1/2} + O\left(L^{-1/2}\right),$$

where $\mu_1 = 0.2865$ and $\sigma_1 = 0.4434$. This means that the fraction of genotypes belonging to a phenotype of log-size $\log N$ is also a normal distribution with the same variance but shifted up to a mean value $\mu_L + \sigma_L^2$ [16]. The two distributions are sketched in figure 11. This picture illustrates why most genotypes (say 99% or 99.9% of them) are only found in a fraction of the largest phenotypes (the orange-colored region in the figure). In particular, it explains why real phenotypes found in nature belong to this top region of the size distribution [13].

We can make this argument more quantitative. Let p be the fraction of genotypes in the blue region of figure 11 (as we said, $p = 0.99$ or $p = 0.999$). If N_p is the lowest size limiting this region from below, then

$$\log N_p = \mu_L + \sigma_L^2 - \sqrt{2}\sigma_L z_p, \qquad p = 1 - \frac{1}{2}\mathrm{erfc}(z_p), \tag{21}$$

$\mathrm{erfc}(z)$ being the complementary error function. From this size we can now compute the fraction of phenotypes to which those genotypes belong (the orange region of figure 11) as

$$q = \frac{1}{2}\mathrm{erfc}(\zeta_p), \qquad \zeta_p = \frac{\log N_p - \mu_L}{\sqrt{2}\sigma_L} = \frac{\sigma_L}{\sqrt{2}} - z_p. \tag{22}$$

This is the fraction of 'visible' phenotypes—those that evolution can find in a random exploration of the genotype space. Table 1 shows that z_p is not very sensitive to the precise value of p, so a reasonable estimate for it is $z_p \approx 2$.

Table 1. Values of z_p for different choices of the fraction p.

p	z_p
0.99	1.644 976
0.999	2.185 124
0.9999	2.629 7417

Now, an asymptotic approximation of $\mathrm{erfc}(u)$ when $u \to \infty$ is [59, equation (7.12.1)]

$$\mathrm{erfc}(u) \sim \frac{e^{-u^2}}{\sqrt{\pi}u}\left[1 + O(u^{-2})\right],$$

therefore, as $L \to \infty$,

$$q \sim \frac{1}{2\sqrt{\pi}} \frac{\sqrt{2}}{\sigma_L - \sqrt{2}z_p} e^{-\left(\frac{\sigma_L}{\sqrt{2}} - z_p\right)^2} \sim \frac{1}{\sqrt{2\pi}\sigma_1 L^{1/2}} e^{-\sigma_1^2 L/2} = \frac{0.9}{L^{1/2}}(1.1)^{-L}.$$

In other words, the fraction of visible phenotypes for a fixed, large value of the fraction of genotypes, *decreases exponentially* with the length L of the genotype. (Notice how this asymptotic estimate *does not* depend on the actual value of z_p.)

But, on the other hand, the total number of phenotypes for RNA secondary structures (with stacks formed by at least two consecutive nucleotide pairs and terminal loops with at least three unpaired nucleotides) grows with L as $1.48L^{-3/2}(1.85)^L$ [14, 60, 61]. Therefore, the absolute number of different, large phenotypes covered by a fraction p of genotypes actually *grows exponentially* with L as $1.33L^{-2}(1.68)^L$.

In summary, the visible phenotypes are only a negligible fraction of all the possible phenotypes that could potentially exist, and nevertheless, the absolute number of them is still huge. So evolution has a lot of variability to choose from even if it only 'sees' a tiny bit of it. But the question remains whether a high replicative ability can compensate for this 'blindness' so as to bring any of these hidden phenotypes to light. In the next subsection we will explain why we think this is highly unlikely to happen.

5.2. Attainable values of phenotype replicative ability

The distribution of replicative abilities of possible phenotypes is mostly unknown, but its range of values can be guessed based on empirical evidence. A paradigmatic example of increase in replicative ability is provided by Spiegelman's experiment where, allowing for arbitrary changes in their length, RNA sequences attained a 15-fold increase in replicative speed [62]. In a more realistic cellular environment, a measure of replicative ability is given by the processivity of RNA Pol II. This protein synthesizes RNA at a speed between 1 kb min^{-1} and 6 kb min^{-1}, with a clear peak around 3 kb min^{-1} and little difference between genes [63]. It seems reasonable to assume that biochemical constraints bound the possible values of the replicative ability of phenotypes, even if other traits are under selection, to a relatively narrow range that spans from a few-fold increase to an order of magnitude.

For the sake of simplicity, let us therefore assume that the replicative ability of a set of phenotypes (for example, those able to accomplish a specific task) follows a Gaussian distribution with average $\langle r \rangle$ and variance σ_r^2. Then, the maximum value of M occurrences follows a peaked distribution around the average value $h_M \sim \langle r \rangle + \sigma_r\sqrt{2\log M}$ [64]. If $M \sim bL^\alpha a^L$,

then $h_M \sim \langle r \rangle + \sigma_r \sqrt{2L \log a}$. Let us now compare the average value of two sets, the first one including all possible phenotypes $M \sim 1.48 L^{-3/2} (1.85)^L$, and the second one embracing those phenotypes covering nearly all genotypes, $M' \sim 1.33 L^{-2} (1.68)^L$, as calculated in the previous section. Thus, $h_M \sim \langle r \rangle + 1.11 \sigma_r L^{1/2}$ and $h_{M'} \sim \langle r \rangle + 1.02 \sigma_r L^{1/2}$. The latter is not even 10% lower than the former. In other words, among the accessible phenotypes the population can find phenotypes with replicative abilities comparable to the largest ones available in the whole phenotype space.

6. Discussion and conclusions

Phenotypes can be described as connected networks of genotypes mutually accessible through mutations. In fixed environments, the fitness of a phenotype corresponds to the largest eigenvalue of the transition matrix associated to the network of genotypes with known replicative abilities. In this contribution we have shown that, in the simplified case where all genotypes in a phenotype have the same replicative ability, the transition between two phenotypes can be successively approximated, with decreasing precision, by the relationship between the two eigenvalues of the phenotypes, the average degree of their genotype networks and finally the log-size of the phenotypes. This latter case is interesting due to the existence of simple computational methods to estimate the size of a phenotype and, especially, because it measures the quantitative relevance of phenotype size in adaptation. In the current context, phenotype size is a measure of entropy and also of robustness of the phenotype [13] and, as such, its turns out to be an essential component of phenotype fitness. Updated representations of fitness landscapes that include the networked nature of phenotypes—such as adaptive multiscapes [22]—become essential to re-educate our intuition on the outcomes of the evolutionary process.

Our approach to the description of the transition between phenotypes has been necessarily simple. We have considered a two-peak landscape for replicative ability and calculated the eigenvector of the joint transition matrix, which represents mutation-selection equilibrium. There are multiple studies that, inspired by the overarching concept of punctuated equilibria [65], have explored the speed of the transition of a population of replicators between two loosely connected networks. In all such studies, sudden transitions in genotype spaces have been identified [12]. Early descriptions of sudden transitions corresponded to adaptation to increasingly fitter phenotypes [29, 66], akin to the scenario explored here. In neutral networks with community structure, the population mostly concentrates in the largest community [45], though sudden transitions occur every time a larger community is found [67]. This phenomenology is analogous to that observed in rough fitness landscapes with complex topology, where the largest fraction of the population is found within a small subset of connected nodes, experiencing sudden shifts in genome space under smooth environmental changes [42], even if phenotypes are not explicitly defined [44]. Numerical simulations of the transition between two phenotypes when the replication rate is smoothly varied (results not shown) yield fast transitions of the type described in previous works. All these observations are well understood in a theoretical framework where networks are visualized as ensembles that compete for resources [43]. Transitions between two such networks can be smooth or sudden, with all possibilities in between, depending on the number of connector links between the networks and the eigenvector centrality of the connecting nodes. Therefore, the strength of the transition and the fraction of population in either network can be tuned through an appropriate election of the nodes than link one network to another. The fact that adaptive transitions in

populations of replicators embedded within a GP map are sudden is consistent with the existence of a reduced number of links between phenotypes and the expectation that most of these connections link peripheral genotypes. The topology of genotype networks, unlike in synthetic examples of neutral networks [67], cannot be modified at will. Community structure seems to be a generic property of realistic GP maps, be these communities mutually neutral, different phenotypes, or a subset of nodes in a fitness landscape. If this is so, sudden transitions in genotype spaces should be the rule also in natural systems (see e.g. [68])—though it might be difficult to disentangle the role played by various variables, such as phenotype size, replicative ability, environmental changes (which may modify at once the two previous variables [22]), or increases in robustness.

Large phenotypes embrace various evolutionary advantages, not all of them adaptive. First, there is a dynamical advantage, known as phenotypic bias, due to the fact that the typical discovery time of a phenotype in blind searches is proportional to the inverse of its frequency [35]. Recent studies have related this phenotype bias to a simplicity bias, arguing that phenotypes with many genotypes (resulting from a GP map) have to be simple in terms of algorithmic information theory and Kolmogorov complexity [69, 70]. This interesting relationship provides an additional way of predicting transitions between phenotypes, where phenotype size would be substituted by phenotype complexity, a quantity also simple to estimate from a computational viewpoint [71]. Second, the average robustness of phenotypes (their average degree) increases with its size as $\langle k \rangle \propto \log N$; that is, the larger the phenotype, the more robust its nodes are. Higher robustness (higher entropy) confers an immediate adaptive advantage. Third, larger phenotypes also have further access to evolutionary novelty, by guaranteeing navigability of the genotype space and facilitating contact with a higher diversity of phenotypes. Altogether, selection of larger (hence fitter) phenotypes appears as an evolutionary trend that could entail a form of irreversibility in evolution. This process is also related to the unfathomable size of genotype spaces: networks of genotypes become so large, even for relatively short sequence lengths, that natural populations are unable to explore any significant portion of them, even in substantial evolutionary time, causing a perpetual drift to more robust regions: they are never stably sitting at the top of a hill.

It has not escaped our notice that the consistent observation that only large phenotypes are found in natural RNA sequences (and probably in any realistic GP map) immediately suggests an alternative interpretation in the light of our results, namely, that the number of phenotypes with significantly larger replicative ability and typical (or smaller) size is actually negligible. In other words, the minimum size of phenotypes found in nature may actually bound the loss in replicative ability, and not the other way round. Should it be otherwise, why would smaller phenotypes with larger λ not be gradually fixed through natural selection? In favor of this alternative view comes the observation of how powerful natural selection is to select for phenotypes that would be never found under blind searches [72], but that can be attained under parsimonious incorporation of increasingly rare (at least at first sight) solutions. Finally, it cannot be discarded that selection for improved replicative ability (or for optimized functionality) and selection for higher robustness (or higher phenotype size) occur concomitantly. If a phenotype highly optimized for function is too rare to guarantee sufficient robustness, the size of such phenotype could be enlarged through modification of traits that preserve function and increase size (though these are usually not included in simple models), such as genotype length [73, 74], the emergence of additional levels in the GP map [15] or the formation of complex interacting molecular ensembles [75]. Nature always finds a way.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Acknowledgments

The authors acknowledge discussions with L F Seoane. This study was funded by Grants PID2020-113284GB-C21 (S M), PGC2018-098186-B-I00 (P C and J A C), PID2022-141802NB-I00 (J A C), PID2019-109320GB-100 and PID2022-142185NB-C22 (P C), and PID2021-122936NB-I00 (J A) all funded by MCIN/AEI/10.13039/501100011033 and by 'ERDF A way of making Europe'.

ORCID iDs

Pablo Catalán https://orcid.org/0000-0003-2826-4684
Juan Antonio García-Martín https://orcid.org/0000-0003-0993-4064
Jacobo Aguirre https://orcid.org/0000-0003-2196-5103
José A Cuesta https://orcid.org/0000-0001-9890-9367
Susanna Manrubia https://orcid.org/0000-0003-0134-2785

References

[1] Stadler P F and Stadler B M R 2006 *Biol. Theory* **1** 268–79
[2] Ahnert S E 2017 *J. R. Soc. Interface* **14** 20170275
[3] Manrubia S *et al* 2021 *Phys. Life Rev.* **38** 55–106
[4] Ogbunugafor C B 2020 *Genetics* **214** 749–54
[5] Louis A A 2016 *Stud. Hist. Phil. Sci.* C **58** 107–16
[6] Bastolla U, Porto M, Roman H E and Vendruscolo M 2003 *J. Mol. Biol.* **56** 243–54
[7] Ciliberti S, Martin O C and Wagner A 2007 *Proc. Natl Acad. Sci. USA* **104** 13595–6
[8] Matias Rodrigues J F and Wagner A 2011 *BMC Syst. Biol.* **5** 39
[9] Schultes E A and Bartel D P 2000 *Science* **289** 448–52
[10] Aguirre J, Buldú J M, Stich M and Manrubia S C 2011 *PLoS One* **6** e26324
[11] Yubero P, Manrubia S and Aguirre J 2017 *Sci. Rep.* **7** 13813
[12] Aguirre J, Catalán P, Cuesta J A and Manrubia S 2018 *Open Biol.* **8** 180069
[13] Dingle K, Schaper S and Louis A A 2015 *Interface Focus* **5** 20150053
[14] Cuesta J A and Manrubia S 2017 *J. Theor. Biol.* **419** 375–82
[15] Catalán P, Wagner A, Manrubia S and Cuesta J A 2018 *J. R. Soc. Interface* **15** 20170516
[16] Garcia-Martin J A, Catalán P, Cuesta J A and Manrubia S 2018 *Europhys. Lett.* **123** 28001
[17] Villanueva A, Secaira-Morocho H, Seoane L F, Lázaro E and Manrubia S 2022 *Biophysica* **2** 381–99
[18] Jörg T, Martin O C and Wagner A 2008 *BMC Bioinform.* **9** 464
[19] Wagner A 2014 *Proc. R. Soc.* B **281** 20132763
[20] Greenbury S F, Schaper S, Ahnert S E and Louis A A 2016 *PLoS Comput. Biol.* **12** e1004773
[21] Greenbury S F, Johnston I G, Louis A A and Ahnert S E 2014 *J. R. Soc. Interface* **11** 20140249
[22] Catalán P, Arias C F, Cuesta J A and Manrubia S 2017 *Biol. Direct* **12** 7
[23] Dall'Olio G M, Bertranpetit J, Wagner A and Laayouni H 2014 *PLoS One* **9** e99424
[24] Wright S 1932 *Proc. 6th Int. Congr. Genet.* vol 1 pp 356–66
[25] Mustonen V and Lässig M 2009 *Trends Genet.* **25** 111–9
[26] Laland K *et al* 2014 *Nature* **514** 161–4
[27] Svensson E I and Calsbeek R 2012 *The Adaptive Landscape in Evolutionary Biology* (Oxford University Press)
[28] Aguirre J 2022 *Nat. Ecol. Evol.* **6** 1599–600

[29] Schuster P and Swetina J 1988 *Bull. Math. Biol.* **50** 635–60
[30] Wolff A and Krug J 2009 *Phys. Biol.* **6** 036007
[31] Khatri B S, McLeish T C B and Sear R P 2009 *Proc. Natl Acad. Sci. USA* **106** 9564–9
[32] Wilke C O, Wang J L, Ofria C, Lenski R E and Adami C 2001 *Nature* **412** 331–3
[33] Codoñer F M, Darós J A, Solé R V and Elena S F 2006 *PLoS Pathog.* **2** e136
[34] Cowperthwaite M C and Meyers L A 2007 *Annu. Rev. Ecol. Syst.* **38** 203–30
[35] Schaper S and Louis A A 2014 *PLoS One* **9** e86635
[36] Catalán P, Manrubia S and Cuesta J A 2020 *J. R. Soc. Interface* **17** 20190843
[37] Fontana W, Konings D A, Stadler P F and Schuster P 1993 *Biopolymers* **33** 1389–404
[38] Ancel L W and Fontana W 2000 *J. Exp. Zool.* **288** 242–83
[39] Schuster P 2006 *Rep. Prog. Phys.* **69** 1419–77
[40] Arias C F, Catalán P, Manrubia S and Cuesta J A 2014 *Sci. Rep.* **4** 7549
[41] Catalán P 2017 Models in molecular evolution: the case of toyLIFE *PhD Thesis* Universidad Carlos III
[42] Aguirre J, Buldú J M and Manrubia S C 2009 *Phys. Rev.* E **80** 066112
[43] Aguirre J, Papo D and Buldú J M 2013 *Nat. Phys.* **9** 230–4
[44] Aguirre J and Manrubia S 2015 *Sci. Rep.* **5** 9664
[45] Capitán J A, Aguirre J and Manrubia S 2015 *Chaos Solitons Fractals* **72** 99–106
[46] van Nimwegen E, Crutchfield J P and Huynen M 1999 *Proc. Natl Acad. Sci. USA* **96** 9716–20
[47] Wagner A 2011 *The Origins of Evolutionary Innovations* (Oxford University Press)
[48] Manrubia S and Cuesta J A 2017 *J. R. Soc. Interface* **14** 20160976
[49] Martin N S and Ahnert S E 2022 *J. R. Soc. Interface* **19** 20220072
[50] Huynen M A 1996 *J. Mol. Evol.* **43** 165–9
[51] Reidys C M, Forst C V and Stadler P F 2001 *Bull. Math. Biol.* **63** 57–94
[52] Reidys C M 2002 *Combinatorial Computational Biology of RNA* (Springer)
[53] Poznanović S and Heitsch C E 2014 *J. Math. Biol.* **69** 1743–72
[54] Li H, Helling R, Tang C and Wingreen N 1996 *Science* **273** 666–9
[55] Kin T, Yamada K, Terai G, Okida H, Yoshinari Y, Ono Y, Kojima A, Kimura Y, Komori T and Asai K 2007 *Nucleic Acids Res.* **35** D145–8
[56] Diener T O 1971 *Virology* **45** 411–28
[57] Flores R, Serra P, Minoia S, Serio F D and Navarro B 2012 *Front. Microbiol.* **3** 217
[58] Catalán P, Elena S F, Cuesta J A and Manrubia S 2019 *Viruses* **11** 425
[59] Olver F W J, Olde Daalhuis A B, Lozier D W, Schneider B I, Boisvert R F, Clark C W, Miller B R, Saunders B V, Cohl H S and McClain M A (eds) 2023 *NIST Digital Library of Mathematical Functions*, *Release 1.1.9 of 2023-03-15* (available at: https://dlmf.nist.gov/)
[60] Stein P and Waterman M 1979 *Discrete Math.* **26** 261–72
[61] Schuster P, Fontana W, Stadler P F and Hofacker I L 1994 *Proc. R. Soc.* B **255** 279–84
[62] Mills D R, Peterson R L and Spiegelman S 1967 *Proc. Natl Acad. Sci. USA* **58** 217–24
[63] Muniz L, Nicolas E and Trouche D 2021 *EMBO J.* **40** e105740
[64] de Haan L and Ferreira A 2006 *Extreme Value Theory: An Introduction* (Springer)
[65] Eldredge N and Gould S J 1972 Punctuated equilibria: an alternative to phyletic gradualism *Models in Paleobiology* ed T J M Schopf (Freeman Cooper) pp 82–115
[66] Huynen M A, Stadler P F and Fontana W 1996 *Proc. Natl Acad. Sci. USA* **93** 397–401
[67] Wilke C O 2001 *Bull. Math. Biol.* **63** 715–30
[68] Koelle K, Cobey S, Grenfell B and Pascual M 2006 *Science* **314** 1898–903
[69] Dingle K, Camargo C Q and Louis A A 2018 *Nat. Commun.* **9** 1–7
[70] Johnston I G, Dingle K, Greenbury S F, Camargo C Q, Jonathan P K D, Ahnert S E and Louis A A 2022 *Proc. Natl Acad. Sci. USA* **119** e2113883119
[71] Dingle K, Novev J K, Ahnert S E and Louis A A 2022 *J. R. Soc. Interface* **19** 20220694
[72] Dawkins R 1996 *Climbing Mount Improbable* (Norton)
[73] Cuypers T D and Hogeweg P 2012 *Genome Biol. Evol.* **4** 212–29
[74] Cuypers T D and Hogeweg P 2014 *PLoS Comput. Biol.* **10** e1003547
[75] Colizzi E S and Hogeweg P 2014 *Genome Biol. Evol.* **6** 1990–2007