Statistical theory of phenotype abundance distributions: A test through exact enumeration of genotype spaces

Juan Antonio García-Martín[1,2,3], Pablo Catalán[1,4], Susanna Manrubia[1,2], and José A. Cuesta[1,4,5,6]
[1]Grupo Interdisciplinar de Sistemas Complejos (GISC), Madrid, Spain
[2]Programa de Biología de Sistemas, Centro Nacional de Biotecnología, CSIC, Madrid, Spain
[3]Bioinformatics for Genomics and Proteomics, Centro Nacional de Biotecnología, CSIC, Madrid, Spain
[4]Departamento de Matemáticas, Universidad Carlos III de Madrid, Leganés, Madrid, Spain
[5]Instituto de Biocomputación y Física de Sistemas Complejos (BIFI), Universidad de Zaragoza, Spain
[6]UC3M-BS Institute of Financial Big Data (IFiBiD), Universidad Carlos III de Madrid, Getafe, Madrid, Spain

The evolutionary dynamics of molecular populations are strongly dependent on the structure of genotype spaces. The map between genotype and phenotype determines how easily genotype spaces can be navigated and the accessibility of evolutionary innovations. In particular, the size of neutral networks corresponding to specific phenotypes and its statistical counterpart, the distribution of phenotype abundance, have been studied through multiple computationally tractable genotype-phenotype maps. In this work, we test a theory that predicts the abundance of a phenotype and the corresponding asymptotic distribution (given the compositional variability of its genotypes) through the exact enumeration of several GP maps. Our theory predicts with high accuracy phenotype abundance, and our results show that, in navigable genotype spaces —characterised by the presence of large neutral networks—, phenotype abundance converges to a log-normal distribution.

It has been suggested that the abundance S_{est} of a phenotype can be estimated as follows [1]. If, for a given phenotype, a variable v_i could measure the average number of different letters of the alphabet that show up at site i of its sequences, then

$$S_{\text{est}} = v_1 v_2 \cdots v_L \tag{1}$$

if the genotype is a chain of length L. Suppose an alphabet of k letters, choose a phenotype and count in how many of its genotypes, $m_{\alpha,i}$, letter α shows up at site i. A suitable definition of the *versatility* of site i is

$$v_i = \frac{1}{M_i}\sum_{\alpha=1}^{k} m_{\alpha,i}, \quad M_i \equiv \max\{m_{1,i},\ldots,m_{k,i}\}. \tag{2}$$

In order to evaluate the goodness of this definition, we have tested it for different GP maps regarding how well it predicts the abundance of a specific phenotype component and its relationship with the distribution of phenotype abundances. First, we have folded all RNA sequences of length $L = 16$ and classified them according to their secondary structures (a proxy for their phenotype). Second, we have analysed a variant of this model, made of RNA sequences containing only two complementary bases, in this case G and C. Third, we have analysed the HP model for lattice proteins, where a protein is represented by a self-avoiding chain of hydrophobic (H) or polar (P) beads on a lattice, in its compact and non-compact versions. Finally, we have also analysed tOYLIFE, a multilevel model of a simplified cellular biology [2] in which binary sequences are first mapped to HP-like proteins that interact between themselves, with the genome, and with metabolites. The phenotype is defined by the set of metabolites that a given sequence is able to catabolise. Figure 1 compares the abundance of phenotypes in exact enumerations of genotype spaces with the prediction of Eqs. (1) and (2). A description of all variants of the former GP maps studied and a full discussion of the results can be found in [3].

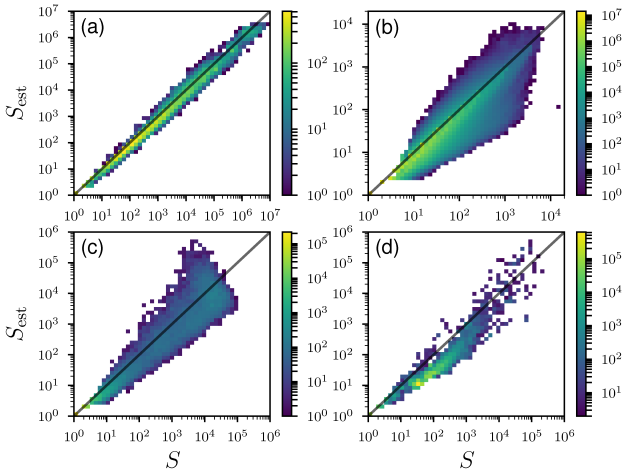

Fig. 1. Log-log-log histograms of the estimated abundance [S_{est} calculated as in Eq. (1)], *versus* actual abundance (S) of the connected components of different GP maps: (a) four-letter RNA of length $L = 16$, (b) two-letter GC-RNA of length $L = 30$, (c) compact HP model 5×6 with $U(HH) = -1$, and (d) tOYLIFE for two genes.

The vastness of genotype spaces prevents a complete characterisation based in computational approaches. Astronomically large numbers are involved in calculations with sequences of length well below that typically found in biochemical processes. An understanding of the structure of realistic GP maps demands further theoretical developments that can be extrapolated to arbitrarily long sequences. The definition of useful quantities such as versatility allows for reliable estimations of the abundance of phenotypes and for the derivation of the expected distribution. Our results yield that distribution in RNA of any length, as well as an estimation of the number of genotypes folding into an arbitrary (typical) structure. Similar derivations should be possible for other GP maps endowed with consistent definitions of phenotype.

[1] S. Manrubia and J. A. Cuesta, Distribution of genotype network sizes in sequence-to-structure genotype-phenotype maps, J. R. Soc. Interface **14**, 20160976 (2017).

[2] C. F. Arias, P. Catalán, S. Manrubia, and J. A. Cuesta, tOYLIFE: a computational framework to study the multi-level organization of the genotype-phenotype map, Sci. Rep. **4**, 7549 (2014).

[3] J. A. García-Martín, P. Catalán, S. Manrubia, and J. A. Cuesta, Statistical theory of phenotype abundance distributions: a test through exact enumeration of genotype spaces, Europhys. Lett. (submitted).