A LETTERS JOURNAL EXPLORING THE FRONTIERS OF PHYSICS

OFFPRINT

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

JUAN ANTONIO GARCÍA-MARTÍN, PABLO CATALÁN, SUSANNA MANRUBIA and JOSÉ A. CUESTA

EPL, **123** (2018) 28001

Please visit the website
www.epljournal.org

A Letters Journal Exploring the Frontiers of Physics

AN INVITATION TO SUBMIT YOUR WORK

epljournal.org

The Editorial Board invites you to submit your letters to EPL

EPL is a leading international journal publishing original, innovative Letters in all areas of physics, ranging from condensed matter topics and interdisciplinary research to astrophysics, geophysics, plasma and fusion sciences, including those with application potential.

The high profile of the journal combined with the excellent scientific quality of the articles ensures that EPL is an essential resource for its worldwide audience. EPL offers authors global visibility and a great opportunity to share their work with others across the whole of the physics community.

Run by active scientists, for scientists

EPL is reviewed by scientists for scientists, to serve and support the international scientific community. The Editorial Board is a team of active research scientists with an expert understanding of the needs of both authors and researchers.

OVER
568,000
full text downloads in 2015

18 DAYS
average accept to online publication in 2015

20,300
citations in 2015

"We greatly appreciate the efficient, professional and rapid processing of our paper by your team."

Cong Lin
Shanghai University

Six good reasons to publish with EPL

We want to work with you to gain recognition for your research through worldwide visibility and high citations. As an EPL author, you will benefit from:

1. **Quality** – The 60+ Co-editors, who are experts in their field, oversee the entire peer-review process, from selection of the referees to making all final acceptance decisions.
2. **Convenience** – Easy to access compilations of recent articles in specific narrow fields available on the website.
3. **Speed of processing** – We aim to provide you with a quick and efficient service; the median time from submission to online publication is under 100 days.
4. **High visibility** – Strong promotion and visibility through material available at over 300 events annually, distributed via e-mail, and targeted mailshot newsletters.
5. **International reach** – Over 3200 institutions have access to EPL, enabling your work to be read by your peers in 100 countries.
6. **Open access** – Articles are offered open access for a one-off author payment; green open access on all others with a 12-month embargo.

Details on preparing, submitting and tracking the progress of your manuscript from submission to acceptance are available on the EPL submission website **epletters.net**.

If you would like further information about our author service or EPL in general, please visit **epljournal.org** or e-mail us at **info@epljournal.org**.

EPL is published in partnership with:

European Physical Society | Società Italiana di Fisica | EDP Sciences | IOP Publishing

A Letters Journal Exploring the Frontiers of Physics

July 2018

EPL, **123** (2018) 28001 www.epljournal.org
doi: 10.1209/0295-5075/123/28001

Focus Article

Statistical theory of phenotype abundance distributions: A test through exact enumeration of genotype spaces[(a)]

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 - Zaragoza, Spain*
[6] *UC3M-BS Institute of Financial Big Data (IFiBiD), Universidad Carlos III de Madrid - Getafe, Madrid, Spain*

received 9 June 2018; accepted in final form 27 July 2018
published online 13 August 2018

PACS 87.10.-e – General theory and mathematical aspects
PACS 87.15.A- – Theory, modeling, and computer simulation
PACS 87.23.Kg – Dynamics of evolution

Abstract – 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.

Introduction. – How the genetic information maps into functional phenotypes (the so-called genotype-to-phenotype, or GP, map) critically conditions the dynamics of evolution [1,2]. Genotypes encode the information to generate phenotypes and in the process of replication undergo all sorts of mutations. The second basic mechanism of evolution, selection, acts upon phenotypes. Standard approaches to evolutionary dynamics have traditionally overlooked the fact that genotype and phenotype are connected through very complex mechanisms, and that the latter may have strong effects on the dynamics.

Genotype spaces can be depicted as networks, with nodes representing genotypes and links joining pairs of genotypes mutually accessible through a mutation. Phenotypes are then subsets of nodes in this network, and the GP map describes their distribution in genotype space. As selection acts on phenotypes, evolution within a connected component of a phenotype is neutral (or nearly so). For this reason, they are referred to in the literature as neutral networks (NNs) [3,4]. A characteristic feature of all known GP maps is the strongly heterogeneous distribution of the abundance (number of nodes) of their NNs [5,6]. A few NNs are huge, typically percolating the whole genotype space, whereas most of them are small. This has evolutionary implications. First of all, the existence of huge NNs endows populations with a high genomic variability without bearing any selective cost. Secondly, most phenotypes are not accessible for entropic reasons [7–9]. Besides, large NNs are so interwoven that virtually any pair of them are connected to each other, thus facilitating the search for phenotypes [10,11]. Under this paradigm, evolution is both robust and innovative.

Given the complexity of GP maps, we need to uncover and characterise as many general features as possible. One of them is the abundance distribution of NNs.

[(a)] Contribution to the Focus Issue *Evolutionary Modeling and Experimental Evolution* edited by José Cuesta, Joachim Krug and Susanna Manrubia.

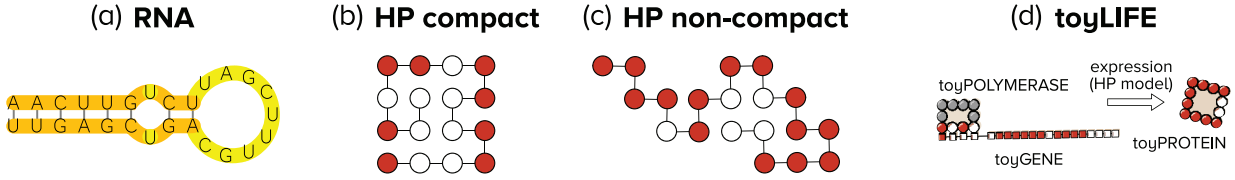

Fig. 1: (Colour online) Schematic representation of the different GP maps exhaustively studied in this work. (a) In RNA, sequences are folded to minimum free energy secondary structures that define the phenotype; (b) in the compact version of the HP model, hydrophobic (H, white circles) and polar (P, red circles) residues adopt the minimum compact energy configuration; (c) in non-compact HP, sequences are assigned to self-avoiding walks of minimum energy; (d) toyLIFE is a multilevel GP map with HP-like sequences that code for compact HP interacting proteins. Phenotype definitions can be found in the Supplementary Material `Supplementarymaterial.pdf` (SM).

The first studies of this distribution often relied on random samplings of the genotype space and considered relatively short RNA molecules [12,13]. These are chains of a two- to four-letter alphabet (A, U, C, G or a subset of those), whose phenotype is identified as a minimum-free-energy folding (secondary structure) [14]. Results pointed to a fat-tailed, decaying distribution [13,15–18] —although whether exponential, power law, or otherwise is far from clear. Later studies of longer molecules (up to 126 letters long) show bell-shaped abundance distributions instead [8].

The first theoretical model addressing this question considered a set of binary sequences with a specific GP mapping rule [19]: the abundance distribution was an unequivocal power law. Later, it was pointed out that two different kinds of distributions —power law and log-normal— are possible [20]. The argument relies on the existence of sites showing low and high compositional variability within a phenotype. Power laws are expected when these positions occupy fixed sites, whereas log-normals arise if their location is not fixed, so that counting the number of arrangements of them in the sequence yields a combinatorial factor. In the case of RNA sequences, low/high variability sites are associated to paired/unpaired nucleotides in the folded structure. A combinatorial calculation of the distribution of paired and unpaired sites can be carried out exactly (see [21] and references therein) and shown to be normal. As the number of low variability sites can be related to the logarithm of the phenotype abundance, the resulting distribution turns out to be log-normal. As a matter of fact, since not only paired sites, but any other structural feature of the folded chain can be shown to have a normal distribution, the argument can be extended even if site variability is affected by other structural elements. The log-normal prediction is thus expected to be quite robust.

Versatility of a site. – An alternative way to look at the problem of estimating phenotype abundance was suggested in the discussion of [20]. 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 the abundance could be estimated as

$$S_{\rm est} = v_1 v_2 \cdots v_L \qquad (1)$$

if the genotype is a chain of length L. This definition is easy to understand if sites are either completely neutral (any mutation maintains the phenotype, $v_i = k$, the size of the alphabet) or fully constrained (any mutation changes phenotype, $v_i = 1$). In a more general case, v_i would take intermediate values.

Given that phenotypes differ in the distributions of their structural motifs, and that the variability of a site is strongly correlated to the motif it sits in, variables v_i can be regarded as phenotype-dependent random variables that take values from a certain distribution. Thus, by the central limit theorem $\ln S$ will be a phenotype-dependent, normally distributed random variable.

Here is a way to estimate one such variable v_i (henceforth referred to as *versatility*) for an alphabet of k letters. We choose a phenotype and count in how many of its genotypes letter α shows up at site i. Let $m_{\alpha,i}$ be that number. Then we define the versatility at site i through

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

The rationale behind this definition relies on assuming that the relative frequencies of each letter of the alphabet at each position i are proportional to the fraction of the space of genotypes associated to the phenotype. It implicitly assumes that the most frequent letter at each position is always characteristic of the phenotype, while other letters, appearing less frequently, may yield sequences corresponding to different phenotypes. For example, if G appears $m_{G,i}$ times and C appears $m_{G,i}/2$ times, other letters being absent, the versatility of that site would be $v_i = 3/2$, meaning that a half of the mutations from G to C at that site change phenotype. When only one letter appears, $v_i = 1$, while $v_i = k$ if all letters are equally likely, recovering the limits of simple models [19,20].

Testing the definition of versatility. – In order to show that the versatility introduced in eq. (2) is a meaningful concept, we have tested it for different GP maps (sketched in fig. 1) 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 $L = 16$, using the Vienna package [22], and classified them

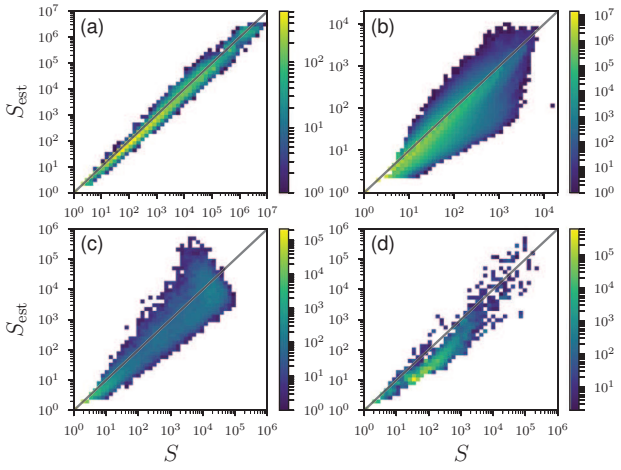

Fig. 2: (Colour online) Log-log-log histograms of the estimated abundance ($S_{\rm est}$ calculated as in (1)), *vs.* actual abundance (S) of the connected components of different GP maps: (a) four-letter RNA of length $L = 16$, (b) two-letter GC-RNA of length $L = 30$, (c) compact HP model 5×6 with $U(HH) = -1$, and (d) tOYLIFE for two genes.

according to their secondary structures. For such a small length, phenotypes are normally fragmented into several connected, neutral components (NCs) of comparable size, but exhaustively folding longer sequences quickly becomes computationally unfeasible. Since NCs behave, to all purposes, as independent NNs, we treat them as independent phenotypes, regardless of whether or not they fold into the same secondary structure. Then, we count how many sequences each NC contains (its abundance, S) and calculate its site versatilities v_i according to the definition (2). The product of them all yields the estimated abundance (1). Figure 2(a) shows a histogram comparing actual and estimated abundances for all the NCs, showing a remarkable agreement. The distinction between NCs and phenotypes becomes less relevant as the length of genotypes grows, as discussed later (see also SM).

A variant of this model is made of RNA sequences containing only two complementary bases, for example G and C (GC-RNA). A two-letter alphabet allows us to study sequences almost twice as long with a similar computational effort [10]. We have repeated the previous analysis for GC-RNA sequences of length $L = 30$, and plotted the result in fig. 2(b). Fragmentation is more frequent in this model, and NCs are generally smaller. This is why their number is so high and why they are so dispersed in fig. 2(b). Also, the largest NCs are three orders of magnitude smaller than those of four-letter RNA sequences. For this model, the versatility of paired sites is strictly 1 because any mutation in such a pair will break the link. Unpaired sites do not have much more freedom either, because a mutation can often create a new link and change the folding. In spite of these constraints, fig. 2(b) shows a clear correlation between S and $S_{\rm est}$, with the overwhelming majority of NCs near the diagonal.

The third GP map that we have analysed is the HP model for lattice proteins [23], where a protein is represented by a self-avoiding chain of hydrophobic (H) or polar (P) beads on a lattice. The energy of a given configuration is calculated from a contact potential,

$$E = \sum_{i<j} U(\sigma_i, \sigma_j) C_{ij}, \tag{3}$$

where $\sigma_i \in \{H, P\}$, $C_{ij} = 1$ when i and j are neighbours on the lattice (with $|i - j| \neq 1$) and $C_{ij} = 0$ otherwise, and $U(\sigma_i, \sigma_j)$ specifies the interaction strength. Several different specific realisations of the model can be found in the literature (see below). For two-dimensional square lattices, compact and non-compact versions of the model have been studied. In compact HP, sequences of length $L = l_1 \times l_2$ are forced to fold into rectangular structures, while non-compact HP considers all self-avoiding walks in the lattice. In fig. 2(c) we show the case example of compact HP 5×6 with a single nonzero energy parameter, $U(H, H) = -1$ where the phenotype is defined as the non-degenerated, minimum energy conformation (see SM).

Finally, we have also analysed tOYLIFE, a multilevel model of a simplified cellular biology [24,25] 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. Consequently, tOYLIFE has a lower genotype level, which translates into proteins (second level), whose interactions add a third, regulatory level. This regulation is altered by the presence of metabolites, which can be catabolised as a result, giving rise to the phenotypic expression at this highest level. Even though the connection between genotype sites and structural elements in this model is far from clear, versatilities can be computed nonetheless. The estimations of phenotype abundances arising from their values, for the case of two genes (length $L = 40$), are compared with actual abundances in fig. 2(d). We can observe a slight but systematic underestimation of abundances. In spite of that, the correlation between S and $S_{\rm est}$ is strong, and the cloud of points runs parallel to the diagonal. The slight underestimation of versatility, however, does not affect the argument leading to the log-normal abundance distribution —only the mean and the variance will be affected.

The prediction of phenotype abundance has been a matter of study, among others due to its relevance for protein designability [26], for molecular robustness and evolvability [27], or in the neutralist-selectionist controversy [8]. Attempts at estimating phenotype abundance have been made using compositional entropy [23,26]. However, the comparison with the predictions obtained through site versatility reveals that versatility has a superior performance (see SM and fig. S1).

Distribution of abundance of RNA NCs. – Figure 3(a) shows the distribution $p(\ln S)$ of the abundance

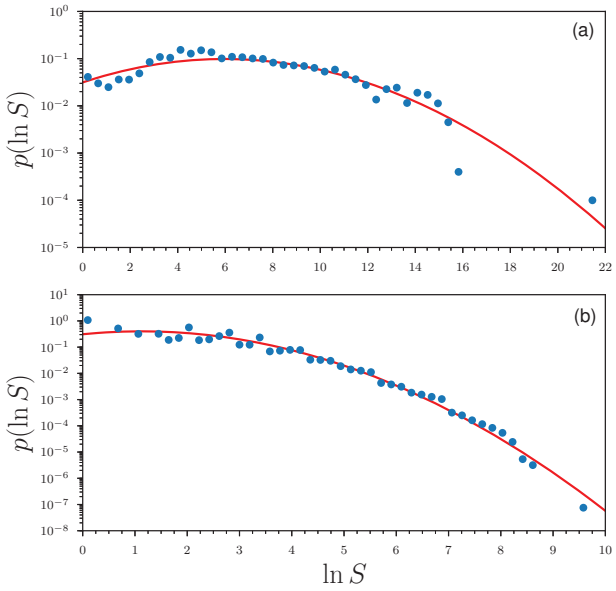

Fig. 3: (Colour online) Log-abundance distributions $p(\ln S)$ for the NCs of (a) four-letter RNA sequences of length $L = 16$ and (b) GC-RNA sequences of length $L = 30$. Dots are the actual values; lines are Gaussian fits to the data.

of RNA sequences of length $L = 16$ in NCs, along with a least-squares fit of the function $\exp[a(\ln S)^2 + b \ln S + c]$, the expected asymptotic distribution according to eq. (1). The length of the sequences is too short to exhibit a perfect Gaussian shape yet: the curve is truncated on the left-hand side and there are deviations for small and large NCs abundances.

Though the abundance distribution of NCs for GC-RNA sequences is a decreasing function with a fat tail (fig. 3(b)), the right tail of a log-normal provides a good fit that captures the slight concavity of the curve. Regardless of the alphabet size, the log-normal distribution is theoretically supported by eq. (1).

The theory developed up to now strictly applies to NCs of phenotypes. However, it was originally inspired by studies reporting a log-normal distribution of *phenotype* abundances [8]. Also, data corresponding to GC-RNA phenotypes compatible with a power-law distribution [16] can be fit at least equally well by a truncated log-normal such as that in fig. 3(b). In the next section we will introduce an effective model that will provide some insights into the specific shapes of these distributions and clarify how the theory asymptotically applies to phenotypes.

Effective two-versatility model for RNA. – Consider long RNA sequences —irrespective of their composition— folded into secondary structures. It has been shown that paired and unpaired sites admit on average a different amount of mutations in a given NC, that is, they differ in neutrality. Asymptotically, the overall neutrality of a phenotype can be well described by two values, each corresponding to one of the structural elements [28,29]. In this vein, we consider now a simplified model with two versatility values: one for paired (v_p) and one for unpaired (v_u) sites (with $1 \leq v_p < v_u \leq k$ for an alphabet of k letters). As neutrality, site versatility depends in principle on many factors other than whether the corresponding base forms a bond. Nevertheless, we do observe that, on average, versatilities associated to paired sites are significantly smaller than those associated to unpaired ones. Interestingly, previous works have identified a clear correlation between RNA secondary structure elements (stems and loops) and nucleotide composition [30,31], giving indirect support to our approximation.

The two-versatility model was introduced [20] to argue for a log-normal distribution of the abundance of RNA sequences in NNs. More precisely, the number of RNA secondary structures with a given number ℓ of paired sites can be shown to be (in the limit $L \to \infty$) proportional to a Gaussian function of ℓ with mean $\mu L - \mu_0$ and standard deviation $\sigma L^{1/2} - \sigma_0 L^{-1/2} + O(L^{-3/2})$ ($\mu = 0.28647$, $\mu_0 = 1.36502$, $\sigma = 0.25510$, $\sigma_0 = 0.00713$). In virtue of (1) and the fact that, within the two-versatility model, $S = v_p^\ell v_u^{L-\ell}$ —hence $\ell \propto \log S$— this immediately leads to a log-normal distribution of S with mean and standard deviation

$$\mu_L = L(\ln v_u - \mu) + \mu_0 + O\left(L^{-1}\right), \quad (4)$$

$$\sigma_L = 2\ln(v_u/v_p)\left(\sigma L^{1/2} - \sigma_0 L^{-1/2}\right) + O\left(L^{-3/2}\right). \quad (5)$$

In order to test this two-versatility model we will use the data of ref. [8] —a collection of estimates of the abundance distribution of RNA secondary structures obtained by sampling random sequences of lengths in the range $L = 20$–126. The resulting distributions are proportional to $Sp(\ln S)$ but, if $p(\ln S)$ is a normal distribution with mean μ_L and standard deviation σ_L, then so is $Sp(\ln S)$, with the same standard deviation but a shifted mean $\mu_L + \sigma_L^2$. Fitting Gaussian functions to these data yields μ_L and σ_L. Then, through eqs. (4), (5) we can infer the corresponding versatilities v_p, v_u —which appear in fig. 4. This plot suggests that these versatilities have well defined asymptotic values for $L \to \infty$, namely $v_p = 1.17 \pm 0.08$, $v_u = 2.79 \pm 0.08$. For comparison, the average versatilities obtained from our data for $L = 16$ are $v_p^{\rm av} = 1.11$, $v_u^{\rm av} = 2.37$.

A caveat is in order here. The results of [8] correspond to the abundance of phenotypes, no matter how many NCs they have, whereas, strictly speaking, the two-versatility model can only be applied to the latter. The surprising agreement of the extrapolated versatilities with those directly obtained from the data for $L = 16$ suggests that for L large, either phenotypes are broken into few NCs, or one of these components is much larger than the others and dominates the abundance of the phenotype. The existence of genetic correlations in NCs seems to cause both effects [6]. Even for short RNA and HP sequences, the largest connected component of a phenotype grows linearly with the abundance of the phenotype, while the number of components either diminishes

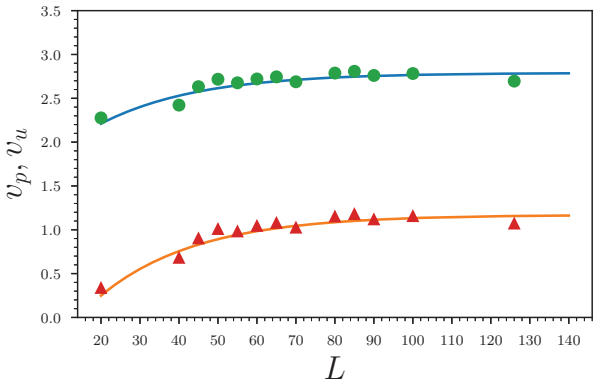

Fig. 4: (Colour online) Average versatilities of unpaired (v_u) and paired (v_p) sites obtained by fitting a two-versatilities model to the sampled abundance distributions of ref. [8] for RNA sequences of different lengths. Lines are fits to data of the form $v_i = v_i^\infty - b_i e^{-c_i L}$, from which the asymptotic values of the two versatilities v_i^∞ are extrapolated.

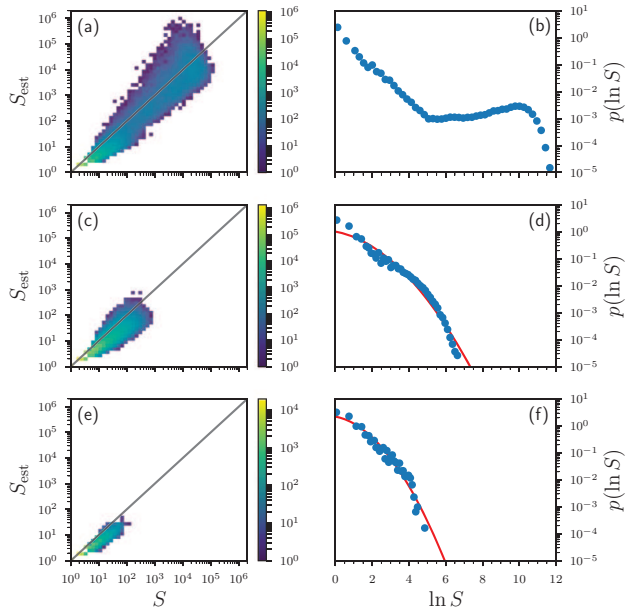

Fig. 5: (Colour online) (a), (c), (e) Log-log-log histograms of the estimated abundance $S_{\rm est}$ *vs.* actual abundance S of the NCs of different HP versions. (b), (d), (f) NCs abundance distributions. (a), (b) Compact HP 5×6 with $U(H,H) = -2.3$ and $U(H,P) = U(P,H) = -1$; (c), (d) non-compact HP30 with $U(H,H) = -1$, and (e), (f) non-compact HP20 $\mathcal{S}$ (based on minimal contact maps) with $U(H,H) = -1$.

with phenotype abundance [10] or remains mostly independent [32]. Therefore, the largest NC becomes more dominant the larger the phenotype, so that the latter is well approximated by a single component. In consequence, the distribution of phenotype abundances is asymptotically equivalent to the distribution of NCs abundances.

The improvement of the fit upon increasing length can be indirectly inferred from the data of ref. [8]. The fits of Gaussian functions to these data are more accurate than the one of fig. 3(a) (see SM and fig. S2), and show that the log-normal behaviour of $p(S)$ is what should be expected for long sequences.

We can apply the two-versatility model to our results with GC-RNA. The effective versatilities are $v_p = 0.75$ and $v_u = 1.32$ (from the data we obtain the exact value $v_p = 1$ and the average $v_u^{\rm av} = 1.43$). As in the case of four-letter RNA (c.f. fig. 4), the values of v_p for short lengths are unphysical ($v_p < 1$). This notwithstanding, effective versatilities are not too far from the average ones, providing an indirect support to the fact that the log-normal distribution for this model has a mean close to 1 —explaining why only the right branch is observed.

Phenotype definition, alphabet size, and navigability of genotype spaces. – Figure 2 suggests that the goodness of the phenotype abundance estimation (1) might depend on the specific GP map. While it works amazingly well for four letter RNA, it is not that good for compact HP or tOyLIFE, which have similarly large NCs. Indeed, high accuracy in that prediction implicitly relies i) on the existence of a clear-cut quantitative relationship between sequence sites and structural elements —which is mediated by a consistent definition of phenotype, and ii) on the presence of a giant NC in phenotypes. The latter seems essential for the abundance of phenotypes to follow a *bona fide* log-normal distribution. Though the relationship between sequence and structure is unequivocal for RNA, it depends on the definition of phenotype in various versions of the HP model (see SM), becomes unavoidably cryptic for tOyLIFE, and might be hard to define in GP maps lacking an easy representation of genotypes as sequences [33]. On the other hand, a comparison of the distribution of abundances in two- and four-letter RNA indicates that the larger the alphabet the larger the components of phenotypes and the better defined the log-normal distributions. These observations are in full agreement with results for the HP model, as shown in the following.

Figure 5 illustrates the performance of versatility and abundance distributions for three additional definitions of phenotype in HP models: compact HP30 with two parameters for energy (fig. 5(a) and (b)), non-compact HP30 ((c) and (d)) and non-compact HP20 with phenotypes defined through *minimal* contact maps ((e) and (f)) that is, the set $\mathcal{S}_{ij}$ formed by those pairs with a nonzero contribution to the folding energy, $\mathcal{S}_{ij} = \{i,j \,|\, U(\sigma_i,\sigma_j)C_{ij} < 0\}$.

Initially, the HP model was implemented in its compact version for computational tractability: notice that the number of different two-dimensional conformations in compact HP30 is 10^8-fold smaller than in non-compact HP30 (table 1). Compact HP versions actually impose unrealistic spatial constraints: two residues i and j can be forced to be in contact without having an associated interaction energy, that is $C_{ij} = 1$, but $U(\sigma_i,\sigma_j) = 0$. Spatial restrictions may therefore assign to a unique phenotype (or NC thereof) sequences whose affiliation easily

Table 1: Data corresponding to the exhaustive enumeration of phenotypes in multiple GP maps. The first column lists the maps studied and some of its quantitative properties: total number of phenotypes, number of non-empty (NE) phenotypes (this quantity resulting from folds with non-negative energy and the large number of degenerated genotypes that are discarded, see SM), number of sequences assigned to a unique phenotype (UaS), average abundance of phenotypes S_{av}, total number of neutral components (NCs), and fraction of non-functional sequences ($f_\emptyset$). Non-compact HP20 (n-c HP20) is included to compare with n-c HP20 with minimal contact maps (n-c HP20 $\mathcal{S}$) as phenotypes (a distribution of phenotype abundances for n-c HP20 can be found in [34]).

Model	Phenotypes	NE phenotypes	UaS	S_{av}	NCs	$f_\emptyset$
RNA30 GC	240944076	432221	1073725603	2484.2	68389814	0.0000151
RNA16 ACGU	5223	648	1712323320	2642474	23092	0.601
compact HP30	13498	13498	187212435	13869.6	362221	0.826
compact HP30[(a)]	13498	13498	258434457	19146.1	1986907	0.759
n-c HP30[(b)]	784924528667	2333498	22466621	9.63	3732449	0.979
n-c HP20	41889578	5310	24900	4.69	6586	0.976
n-c HP20 $\mathcal{S}$	910971	54818	292732	5.34	62379	0.721
t$_{\bigcirc}$Y**LIFE**	$2^{214} \simeq 2.63 \times 10^{64}$	775	134400450	173419.9	1523544	0.9999

[(a)] Data obtained with two energy parameters, $U(H,H) = -2.3$ and $U(H,P) = U(P,H) = -1$.
[(b)] Data from [35].

changes under more natural phenotype definitions [36]. This has an immediate effect on abundance distributions, as fig. 5(b) shows: besides a decrease at small NC sizes, the distribution develops a bump at high abundances. The non-compact versions of HP are difficult to explore exhaustively due to the astronomically large number of possible phenotypes [35]. Still, phenotypes are free from spatial constraints and, as a result, abundance distributions can be fit with a log-normal function (fig. 5(d), (f)). These distributions are qualitatively similar to that obtained for GC-RNA, though NCs are significantly larger in the latter. Smaller NCs could be expected if, instead of the Vienna Package to fold RNA sequences, a model with few energy parameters (such as, *e.g.*, Nussinov algorithm for loop matching [37]) is used.

In either compact or non-compact realisations, folding is calculated by using one [35] or two [23] nonzero energy parameters, examples being $U(H,H) = -1$, as in fig. 2(c)) or $U(H,H) = -2.3$, and $U(H,P) = -1$, *e.g.*, as in fig. 5(a)). Genotypes in these HP models can typically be mapped to more than one phenotype. Traditionally, these degenerated genotypes are discarded, since they have been interpreted as the analogues of intrinsically disordered proteins, and therefore devoid of function. This convention results in one of the most concerning features of classical HP models [38], where an astonishingly large fraction of sequences are systematically not assigned to phenotypes, yielding empty phenotypes and many small and highly fragmented ones (see table 1 for representative examples). It is important to remark that a high fraction of non-functional sequences does not necessarily imply that phenotypes are small and isolated, since other models —where the small fraction of functional sequences is not due to degeneration— do have large and easily navigable phenotypes [24,39,40].

Adding more energy parameters serves to disambiguate the assignation of genotypes to phenotypes, though the increase in the fraction of sequences assigned to phenotypes is however minor (compare the two compact HP30 versions in table 1). Phenotypes defined through contact maps are closer analogues of RNA secondary structure (as in our example with non-compact HP20): contact maps appear as a more natural definition of phenotype that furthermore reduces about 40-fold the number of different phenotypes and notably decreases sequence degeneration (table 1). Also, degeneration diminishes significantly when the size of the alphabet grows. In a systematic study with sequences of length $L = 25$, degeneration is halved when going from two- to four-letter alphabets, and it reaches a few percent for 20-letter representations [41]. Concomitantly, phenotypes become larger and more connected.

The fact that most phenotypes are small, weakly connected and even difficult to navigate in classical HP models [35] raises doubts on their relevance for evolutionary dynamics, speaking in favour of more complex but also more realistic scenarios [38], and certainly supporting non-compact versions of lattice protein models [36]. In agreement with the above, the definition of phenotype critically affects the distribution of abundances, which changes from decreasing functions for two-letter alphabets (as in fig. 5) to functions with a maximum and a fat tail for 20-letter, compact versions [38,42]. Independent studies suggest that minimal alphabets are not optimal in an evolutionary sense [43], further supporting the limited applicability of two-letter models, especially to draw conclusions on evolutionary dynamics. Unfortunately, an exhaustive study of non-compact lattice protein models with more than two letters is, as of today, computationally unfeasible.

Conclusions. – The vastness of genotype spaces prevents a complete characterisation based on computational approaches. A look at table 1 suffices to illustrate the astronomically large numbers involved in calculations with sequences of length well below that typically found in

biochemical processes. The data generated to analyse the different models in this contribution reaches 0.5 TB and, as their diversity shows, would be of limited use in the absence of an accompanying theory. Therefore, an understanding of the structure of realistic GP maps demands further theoretical developments that can be extrapolated to arbitrarily long sequences. We have shown that 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. The knowledge of the asymptotic values v_p and v_u yields 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 the phenotype.

* * *

K. Dingle, E. Ferrada, Ch. Holzgräfe, A. Irbäck and A. Louis are gratefully thanked for sharing their data with us. We acknowledge financial support by the Spanish Ministerio de Economía y Competitividad and FEDER funds of the EU through grants VARIANCE (FIS2015-64349-P; PC, JAC) and ViralESS (FIS2014-57686-P; JAGM, SM).

REFERENCES

[1] Cowperthwaite M. C. and Meyers L. A., *Annu. Rev. Ecol. Evol. Sys.*, **38** (2007) 203.
[2] Aguirre J., Catalán P., Cuesta J. A. and Manrubia S., *Open Biol.*, **8** (2018) 180069.
[3] Schuster P., Fontana W., Stadler P. F. and Hofacker I. L., *Proc. R. Soc. London B*, **255** (1994) 279.
[4] Bornberg-Bauer E., *Biophys. J.*, **73** (1997) 2393.
[5] Wagner A., *The Origins of Evolutionary Innovations* (Oxford University Press) 2011.
[6] Ahnert S. E., *J. R. Soc. Interface*, **14** (2017) 20170275.
[7] Schaper S. and Louis A. A., *PLoS ONE*, **9** (2014) e86635.
[8] Dingle K., Schaper S. and Louis A. A., *J. R. Soc. Interface*, **5** (2015) 20150053.
[9] Catalán P., Arias C. F., Cuesta J. A. and Manrubia S., *Biol. Direct*, **12** (2017) 7.
[10] Grüner W., Giegerich R., Strothmann D., Reidys C., Weber J., Hofacker I. L., Stadler P. F. and Schuster P., *Monatsh. Chem.*, **127** (1996) 375.
[11] Fontana W. and Schuster P., *J. Theor. Biol.*, **194** (1998) 491.
[12] Schuster P. and Stadler P. F., *Comput. Chem.*, **18** (1994) 295.
[13] Grüner W., Giegerich R., Strothmann D., Reidys C., Weber J., Hofacker I. L., Stadler P. F. and Schuster P., *Monatsh. Chem.*, **127** (1996) 355.
[14] Schuster P., *Rep. Prog. Phys.*, **69** (2006) 1419.
[15] Stich M., Briones C. and Manrubia S. C., *J. Theor. Biol.*, **252** (2008) 750.
[16] Ferrada E. and Wagner A., *Biophys. J.*, **102** (2012) 1916.
[17] Cowperthwaite M. C., Economo E. P., Harcombe W. R., Miller E. L. and Meyers L. A., *PLoS Comput. Biol.*, **4** (2008) e1000110.
[18] Aguirre J., Buldú J. M., Stich M. and Manrubia S. C., *PLoS ONE*, **6** (2011) e26324.
[19] Greenbury S. and Ahnert S., *J. R. Soc. Interface*, **12** (2015) 20150724.
[20] Manrubia S. and Cuesta J. A., *J. R. Soc. Interface*, **14** (2017) 20160976.
[21] Cuesta J. A. and Manrubia S., *J. Theor. Biol.*, **419** (2017) 375.
[22] Lorenz R., Bernhart S. H., Höner zu Siederdissen C., Tafer H., Flamm C., Stadler P. F. and Hofacker I. L., *Algorithms Mol. Biol.*, **6** (2011) 26.
[23] Li H., Helling R., Tang C. and Wingreen N., *Science*, **273** (1996) 666.
[24] Arias C. F., Catalán P., Manrubia S. and Cuesta J. A., *Sci. Rep.*, **4** (2014) 7549.
[25] Catalán P., Wagner A., Manrubia S. and Cuesta J. A., *J. Roy. Soc. Interface*, **15** (2018) 20170516.
[26] Larson S. M., England J. L., Desjarlais J. R. and Pande V. S., *Protein Sci.*, **11** (2002) 2804.
[27] Jörg T., Martin O. C. and Wagner A., *BMC Bioinf.*, **9** (2008) 464.
[28] Huynen M. A., *J. Mol. Evol.*, **43** (1996) 165.
[29] Reidys C. M., Forst C. V. and Stadler P. F., *Bull. Math. Biol.*, **63** (2001) 57.
[30] Schultes E. A., Hraber P. T. and LaBean T. H., *RNA*, **3** (1997) 792.
[31] Smit S., Yarus M. and Knight R., *Bioinformatics*, **12** (2006) 1.
[32] Greenbury S. F., Schaper S., Ahnert S. E. and Louis A. A., *PLoS Comput. Biol.*, **12** (2016) e1004773.
[33] Ibáñez-Marcelo E. and Alarcón T., *J. Theor. Biol.*, **356** (2104) 144.
[34] Shahrezaei V., Hamedani N. and Ejtehadi M. R., *Phys. Rev. E*, **60** (1999) 4629.
[35] Holzgräfe C., Irbäck A. and Troein C., *J. Chem. Phys.*, **135** (2011) 195101.
[36] Irbäck A. and Troein C., *J. Biol. Phys.*, **28** (2002) 1.
[37] Nussinov R., Pieczenik G., Griggs J. R. and Kleitman D. J., *SIAM J. Appl. Math.*, **35** (1978) 68.
[38] Buchler N. E. G. and Goldstein R. A., *Proteins*, **34** (1999) 113.
[39] Ciliberti S., Martin O. C. and Wagner A., *Proc. Natl. Acad. Sci. U.S.A.*, **104** (2007) 13595.
[40] Matias Rodrigues J. F. and Wagner A., *PLoS Comput. Biol.*, **5** (2009) e1000613.
[41] Buchler N. E. G. and Goldstein R. A., *J. Chem. Phys.*, **112** (2000) 2533.
[42] Li H., Tang C. and Wingreen N. S., *Proteins*, **49** (2002) 403.
[43] Gardner P. P., Holland B. R., Moulton V., Hendy M. and Penny D., *Proc. R. Soc. London B*, **270** (2003) 1177.