INTERFACE

rsif.royalsocietypublishing.org

Research

Cite this article: Catalán P, Wagner A, Manrubia S, Cuesta JA. 2018 Adding levels of complexity enhances robustness and evolvability in a multilevel genotype–phenotype map. *J. R. Soc. Interface* **15**: 20170516.
http://dx.doi.org/10.1098/rsif.2017.0516

Received: 17 July 2017
Accepted: 1 December 2017

Subject Category:
Life Sciences–Mathematics interface

Subject Areas:
evolution, biomathematics, computational biology

Keywords:
genotype–phenotype map, robustness, evolvability, toyLIFE, HP protein folding, multilevel phenotype

Author for correspondence:
Pablo Catalán
e-mail: pablocatalanfdez@gmail.com

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.3950983.v1.

THE ROYAL SOCIETY PUBLISHING

Adding levels of complexity enhances robustness and evolvability in a multilevel genotype–phenotype map

Pablo Catalán[1,2], Andreas Wagner[3,4,5], Susanna Manrubia[1,6] and José A. Cuesta[1,2,7,8]

[1]Grupo Interdisciplinar de Sistemas Complejos (GISC), Madrid, Spain
[2]Departamento de Matematicas, Universidad Carlos III de Madrid, Madrid, Spain
[3]Department of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland
[4]Santa Fe Institute, Santa Fe, NM, USA
[5]Swiss Institute of Bioinformatics, Zurich, Switzerland
[6]Programa de Biología de Sistemas, Centro Nacional de Biotecnologia, Madrid, Spain
[7]Instituto de Biocomputación y Física de Sistemas Complejos (BIFI), Universidad de Zaragoza, Zaragoza, Spain
[8]Institute of Financial Big Data (IFiBiD), Universidad Carlos III de Madrid, UC3M-BS, Madrid, Spain

PC, 0000-0003-2826-4684; AW, 0000-0003-4299-3840; SM, 0000-0003-0134-2785

Robustness and evolvability are the main properties that account for the stability and accessibility of phenotypes. They have been studied in a number of computational genotype–phenotype maps. In this paper, we study a metabolic genotype–phenotype map defined in toyLIFE, a multilevel computational model that represents a simplified cellular biology. toyLIFE includes several levels of phenotypic expression, from proteins to regulatory networks to metabolism. Our results show that toyLIFE shares many similarities with other seemingly unrelated computational genotype–phenotype maps. Thus, toyLIFE shows a high degeneracy in the mapping from genotypes to phenotypes, as well as a highly skewed distribution of phenotypic abundances. The neutral networks associated with abundant phenotypes are highly navigable, and common phenotypes are close to each other in genotype space. All of these properties are remarkable, as toyLIFE is built on a version of the HP protein-folding model that is neither robust nor evolvable: phenotypes cannot be mutually accessed through point mutations. In addition, both robustness and evolvability increase with the number of genes in a genotype. Therefore, our results suggest that adding levels of complexity to the mapping of genotypes to phenotypes and increasing genome size enhances both these properties.

1. Introduction

Classical evolutionary models do not account for the robustness and evolvability of phenotypes [1]. They thus fail to explain some evolutionary phenomena, such as punctuated equilibria [2,3], constrained evolution [4] or the origins of novelty [5,6]. In recent years, several research groups have tried to understand this important question by studying computational mappings of molecular genotypes to phenotypes. Some of these maps try to remain faithful to biological phenomena, such as RNA secondary structure [7–14], protein secondary structure [15–20], gene regulatory networks [21–24] and metabolic networks [25–28]. More abstract models have also been developed, such as the polyomino [29–31] and toyLIFE [32], as well as the Fibonacci map [33] and simple combinatorial maps [34,35]. The robustness and evolvability of phenotypes have also been recently studied for the artificial life AVIDA system [36].

Even though all these models focus on different aspects of molecular biology, all of them share some common properties. First, the mapping from genotype to phenotype is highly degenerate: many genotypes encode the same phenotype. Additionally, phenotype abundance (the number of genotypes encoding it) is

not evenly distributed: most phenotypes are rare, while a few of them are extremely abundant. The probability density function associated with phenotype abundance is a log-normal distribution for a wide variety of models [14,34], although in some cases it has been found to be a power law [33,34]. This implies that rare phenotypes will not play a central role in evolution: they are hard to find in a genotype space that is filled with abundant phenotypes [12,14,37]. This degeneracy is usually accompanied by the formation of neutral networks, networks of genotypes encoding the same phenotype, in which two genotypes are connected if they differ in one point mutation—i.e. if the strings representing each genotype differ in one letter [7,16,21,25,38]. The degree of a node in such a neutral network—the number of neutral neighbours it has—is called genotypic robustness. It is usually normalized by the total number of neighbours in genotype space, and thus represents the fraction of neighbours with the same phenotype [5]. Inside a particular neutral network, the degree distribution is highly heterogeneous, but usually unimodal [13,39]. Additionally, in RNA, the average degree of a neutral network grows linearly with the logarithm of its abundance [13]—this positive correlation is also observed in the polyomino model [30], the HP model [31] and simple genotype–phenotype maps [34], and is suggested by empirical data [40]. Neutral networks of abundant phenotypes percolate genotype space: they contain genotypes that share almost no letters [5,7,21]. Conversely, most abundant phenotypes are easily accessed from each other: traversing a neutral network, many phenotypes can be found at its boundary [5,21,25]. Moreover, abundant phenotypes are typically found just a few mutations away from a random genotype [5,9,30]—i.e. they are highly evolvable. This means that these phenotypes are easily accessible from any other phenotype, so that the search for new phenotypes among abundant ones is a fast evolutionary process.

In [32], we presented toyLIFE, a multilevel model of a genotype–phenotype map (see figure 1; electronic supplementary material, §S1 for a summary of toyLIFE's definition). toyLIFE includes genes, proteins, regulatory networks and a simple metabolism. Genes are binary strings of fixed length, divided in two regions—a promoter and a coding region (figure 1*a*). The coding region is translated into a sequence of 16 amino-acids that folds on a 4×4 lattice, following the rules of the HP protein-folding model [15–17] (figure 1*a*,*d*). Once folded, we only distinguish proteins by their perimeter and folding energy—note that this definition of folded protein is different from other versions of the HP model. In toyLIFE, there are 2710 different proteins (electronic supplementary material, figure S3), most of them obtained from more than one sequence. However, there are no neutral mutations: every change in a coding region will result (most of the time) in a non-folding protein or (more rarely) in a different functional protein. This is very different from what has been observed in other versions of the HP model [17–20]. Proteins in toyLIFE interact with each other to form dimers, and both proteins and dimers regulate gene expression (figure 1*c*) and interact with metabolites (figure 1*b*). The phenotype in toyLIFE can be defined in multiple ways. Here, we will focus on a metabolic definition of phenotype (figure 1*e*), similar to the one presented in [25,41]: the set of metabolites that a genotype can metabolize (electronic supplementary material, §S2). toyLIFE is, to our knowledge, the only multilevel genotype–phenotype map incorporating genetic dynamics, protein folding, regulatory networks and metabolism.

In this paper, we investigate the characteristics of the metabolic genotype–phenotype map of toyLIFE. First, we want to assess if toyLIFE shares the properties of most computational genotype–phenotype maps studied before. We will see that, in spite of toyLIFE's complexity, its properties are very similar to many of these maps. Second, we wish to explore the robustness and evolvability of this map. The regulatory and metabolic functions of toyLIFE are built on a non-robust, non-evolvable version of the HP model, in which proteins can hardly evolve without going through non-folding intermediate steps. We show in this article that both robustness and evolvability are enhanced by the superposition of additional levels of organization. Third, we will explore how robustness and evolvability change when genome size is increased. We show that they increase significantly, and that we can explain this tendency in the light of toyLIFE's details.

2. Degeneracy of the genotype–phenotype map

The size of genotype space in toyLIFE grows very quickly with the number of genes in a genotype. Since a gene in toyLIFE is a binary string of length 20, there are 2^{20} different genes. A genotype is formed by choosing g genes from this set with replacement (the order of genes is irrelevant). Hence, the number of genotypes with g genes is $\binom{g+2^{20}-1}{g} \approx 10^{6g}/g!$ (electronic supplementary material, §S1). For $g = 2$, this number is 5.5×10^{11}, for $g = 3$, it is 1.9×10^{17}, and for larger values of g it keeps growing almost exponentially. A complete exploration of these genotype spaces is well beyond our computational possibilities in general. However, using computational tricks, we have exhaustively analysed the $g = 2$ and $g = 3$ cases—i.e. we have limited our study to two-gene and three-gene genotypes.

We have restricted ourselves to studying those genotypes that are able to catabolize at least one metabolite—these will be called viable genotypes. The remaining (non-viable) genotypes are unable to catabolize any metabolite. For $g = 2$, there are 1.1×10^9 viable genotypes, representing little more than 0.2% of all genotypes. For $g = 3$, this number is 1.0×10^{15}—approximately 0.5% of all genotypes. In both cases, the great majority of genotypes are unable to catabolize any metabolite. But note that the space of viable genotypes is still enormous.

Among these viable genotypes, many of them catabolize exactly the same metabolites—they encode the same metabolic phenotype. For $g = 2$, there are only 775 different phenotypes, corresponding to an average of 1.4×10^6 genotypes per phenotype. For $g = 3$, there are 26 492 phenotypes, corresponding to an average of 3.8×10^{10} genotypes per phenotype. In other words, for both $g = 2$ and $g = 3$, the degeneracy of the genotype–phenotype map is huge. From now on, we will refer to the set of phenotypes found for $g = 2$ and $g = 3$ as $\mathcal{P}_2$ and $\mathcal{P}_3$, respectively.

The distribution of phenotype abundances is highly skewed (figure 2*a*), similarly to what has been observed for other genotype–phenotype maps. In both cases, the distribution can be empirically fitted to a log-normal distribution. We obtained the parameters empirically from the log-transformed data, using maximum-likelihood. For both $g = 2$ and $g = 3$, the rank distributions (electronic supplementary material, figure S9) show a long tail, confirming that, indeed,

Figure 1. Brief overview of toyLIFE. (*a*) The three basic building blocks of toyLIFE are toyNucleotides, toyAminoacids and toySugars. They can be hydrophobic (H, white) or polar (P, red), and their random polymers constitute toyGenes, toyProteins and toyMetabolites. The toyPolymerase is a special polymer that will have specific regulatory functions. These polymers will interact between each other following an extension of the HP model, for which we have chosen the interaction energies $E_{HH} = -2$, $E_{HP} = -0.3$ and $E_{PP} = 0$ [17]. (*b*) Possible interactions between pairs of toyLIFE elements. toyGenes interact through their promoter region with toyProteins (including the toyPolymerase and toyDimers); toyProteins can bind to form toyDimers, and interact with the toyPolymerase when bound to a promoter; both toyProteins and toyDimers can bind a toyMetabolite at arbitrary regions along its sequence. (*c*) A toyGene is expressed (translated) when the toyPolymerase binds to its promoter region. The sequence of Ps and Hs of the toyProtein will be exactly the same as that of the toyGene coding region. If a toyProtein binds to the promoter region of a toyGene with a lower energy than the toyPolymerase does, it will displace the latter, and the toyGene will not be expressed. This toyProtein acts as an *inhibitor*. The toyPolymerase does not bind to every promoter region. Thus, not all toyGenes are expressed constitutively. However, some toyProteins will be able to bind to these promoter regions. If, once bound to the promoter, they bind to the toyPolymerase with their rightmost side, the toyGene will be expressed, and these toyProteins act as *activators*. (*d*) toyProteins fold on a 4×4 lattice, following a self-avoiding walk (SAW). For each binary sequence of length 16, we fold it into every SAW and compute its folding energy, following the HP model. Then we choose the SAW that yields the minimum folding energy. (*e*) Metabolism in toyLIFE. A toyDimer is bound to a toyMetabolite when a new toyProtein comes in. If the new toyProtein binds to one of the two units of the toyDimer, forming a new toyDimer energetically more stable than the old one, the two toyProteins will unbind and break the toyMetabolite up into two pieces. We say that the toyMetabolite has been catabolized. (Online version in colour.)

while few phenotypes are very abundant, most of them are rare. For $g = 3$, this is especially striking, since 300 phenotypes in $\mathcal{P}_3$ represent nearly 99% of all genotypes—which means that the remaining 1% is distributed among approximately 26 000 phenotypes.

All phenotypes in $\mathcal{P}_2$ are also found in $\mathcal{P}_3$: we can always add a gene whose product does not fold into any protein to a viable two-gene genotype. A pertinent question, therefore, is how abundant the phenotypes belonging to $\mathcal{P}_2$ are in three-gene genotype space. We find that phenotypes in $\mathcal{P}_2$ take up 99.6% of genotypes in $g = 3$ (electronic supplementary material, §S4). This means that the 775 phenotypes in $\mathcal{P}_2$ dominate the space of phenotypes for $g = 3$. Only special combinations of three proteins and three promoters will yield most of the phenotypic diversity observed for $g = 3$. The majority of genotypes will be extensions of two-gene genotypes with a third gene that does not interfere with their function.

We can re-compute the histogram in figure 2*a*(ii) taking the 775 phenotypes from $\mathcal{P}_2$ as a separate set from the remaining 25 717 phenotypes in $\mathcal{P}_3$ that are not in $\mathcal{P}_2$ (electronic supplementary material, §S4). This reveals that the small bump observed in the right part of the distribution of phenotypes in $\mathcal{P}_3$ (figure 2*a*) is due to the phenotypes in $\mathcal{P}_2$. When we eliminate these phenotypes, the resulting distribution is much closer to a log-normal. In a sense, it is as if both sets were somehow independent: one is formed by two-gene genotypes with a third, non-interfering gene, and the other is formed by all combinations of three genes that encode new phenotypes. This influence of $\mathcal{P}_2$ phenotypes decays linearly when genotype size increases (electronic

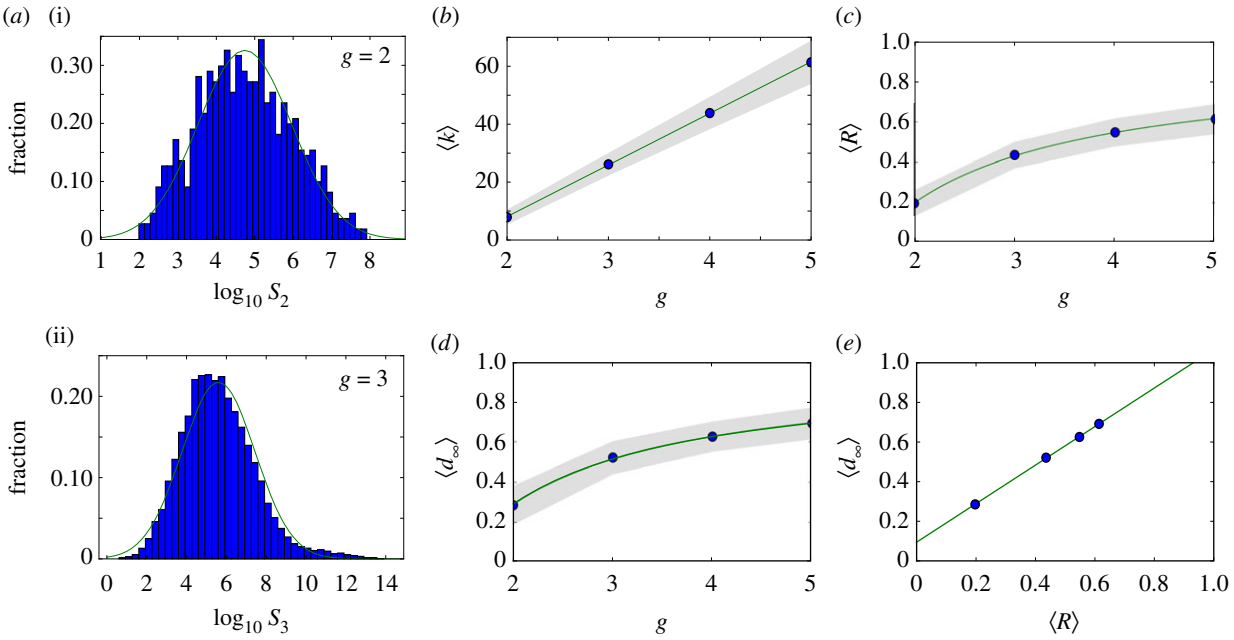


Figure 2. Neutral networks in toyLIFE. (*a*) The distribution of the number of genotypes per phenotype—phenotype abundance, S—for $g = 2$ (S_2, *a*(i)) follows a log-normal distribution, with probability density function $f(x) = (x\sigma\sqrt{2\pi})^{-1} \exp(-(\log x - \mu)^2/2\sigma^2)$, where μ is the mean and σ is the standard deviation of the normally distributed logarithm of the variable. Here $\mu = 4.742$ and $\sigma = 1.224$, obtained using maximum-likelihood. For $g = 3$, the distribution of phenotype abundances (S_3, *a*(ii)) is again very close to a log-normal distribution with $\mu = 5.604$ and $\sigma = 1.838$. The log-normal fit is worse than in (*a*) because there is a small bump in the right part of the distribution, where more abundant phenotypes are—due to the over representation of two-gene phenotypes (see text). (*b*) Average degree of nodes (circles) in neutral networks (see electronic supplementary material, figure S12 for the degree distribution) versus gene number g. The average degree $\langle k \rangle$ of a node grows linearly with gene number g, as $\langle k \rangle = -27.6 + 17.8\,g$ (line). (*c*) Average robustness (circles) versus gene number g. Robustness grows with gene number, and we can find a nonlinear relationship between both variables: $\langle R \rangle = 0.895 - 1.392/g$ (line). (*d*) There is a nonlinear relationship between g and $\langle d_\infty \rangle$, the final distance that is reached in a random walk in which genotypes are forced to get away from the starting genotype every step: $\langle d_\infty \rangle = 0.965 - 1.354/g$ (line). The circles represent $\langle d_\infty \rangle$. (*e*) There is a linear relationship between $\langle d_\infty \rangle$ and the average robustness of the genotypes as obtained in *c*, given by: $\langle d_\infty \rangle = 0.094 + 0.972\langle R \rangle$ (line), very close to the $\langle d_\infty \rangle = \langle R \rangle$ fit. In all cases, the grey area encompasses two standard deviations, and the fits in (*b*–*e*) were obtained using the least-squares method. (Online version in colour.)

supplementary material, §S5), although they keep representing more than 80% of genotype space for $g \leq 13$.

3. Neutral networks in toyLIFE

Robustness can be defined as $R = k/k_{\max}$, where k is the degree of a node in a neutral network, and $k_{\max} = 20g$ is the maximum number of point-mutation neighbours. In other words, R is the normalized degree of a node. We can sample genotypes for different genotype sizes, represented by g (gene number), and plot the histogram of values of R (electronic supplementary material, figure S12). All the resulting distributions are unimodal, as has been observed in other genotype–phenotype maps [5,13]. toyLIFE genotypes become more robust as g increases. In fact, there is a linear relationship between g and $\langle k \rangle$, the average degree of a node in a neutral network (figure 2*b*): $\langle k \rangle = -27.6 + 17.8g$. But $\langle R \rangle = \langle k \rangle/20g$, so we obtain $\langle R \rangle \sim 0.891 - 1.378/g$, which is very close to the least-squares fit $\langle R \rangle = 0.895 - 1.392/g$, shown in figure 2*c*. The linear relationship between $\langle k \rangle$ and g with slope 17.8 indicates that, on average, for every gene we add to a genotype, nearly 18 out of 20 new mutations will be neutral. This implies that $\langle R \rangle$ saturates at a value close to 0.9 when g increases. This result is consistent with the results of §2, which showed that newly added genes rarely interfere with an existing phenotype. These new genes can be viewed as 'junk' in the sense that they do not have any effect on metabolic function and that mutations in their sequence tend to be neutral. We will see later on that junk genes are however important, in that they enhance evolvability in toyLIFE genotypes.

Also, taking into account that $\mathcal{P}_2$ phenotypes dominate in $\mathcal{P}_g$ for $g \leq 13$ (electronic supplementary material, §S5), we can estimate that $S_g \sim 17.8^g S_2$, so $\log S_g \sim q + g\log 17.8$, where q is a constant. Combining this result with the linear relationship between g and $\langle k \rangle$, we obtain for toyLIFE the linear relationship between $\langle k \rangle$ and $\log S$, that has been observed previously for other models [13,30,31,34] (but see figure 3 for a direct verification of this relationship).

In other genotype–phenotype maps, neutral networks tend to have one giant component [18], although this is not always the case: too short RNA sequences form neutral networks that are highly disconnected [13]. Although network analysis is almost impossible for $g \geq 3$, as networks are enormous, for $g = 2$ we can perform network analyses on all 775 phenotypes exhaustively, and compute their connected components (electronic supplementary material, §S7). We observe that most phenotypes are distributed in highly fragmented neutral networks: the genotypes encoding a given phenotype form many disjoint connected components, which are typically small. Abundant phenotypes tend to have a larger number of connected components, and we can find a relatively good power-law fit between the abundance of the phenotype S_2 and the number of components C: $C = 0.25S_2^{0.7}$ (electronic supplementary material, §S7). We also observe a huge variation in the size of connected components in $g = 2$: although more than 98% of connected components are smaller than 1000 nodes, some of them reach up to approximately 10^7 nodes. The high

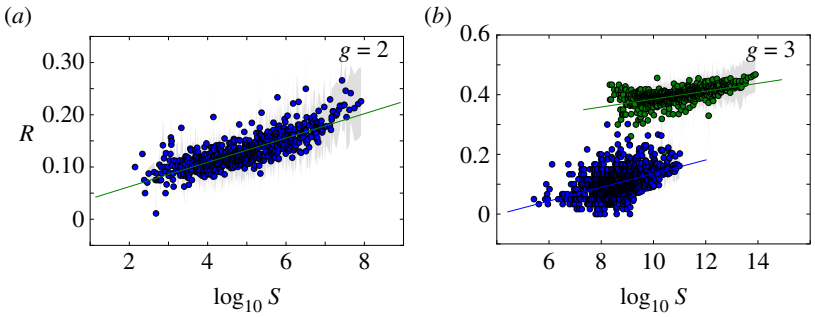


Figure 3. Phenotypic robustness is linearly related to the logarithm of phenotype abundance. For $g = 2$ and $g = 3$, we sampled 10^7 genotypes and computed their robustness. Then we assigned each of them to their corresponding phenotypes, and estimated phenotypic robustness, the average robustness for all genotypes encoding a given phenotype (see text). (*a*) Phenotypic robustness in $g = 2$ versus the logarithm of phenotype abundance. The line represents the power-law relationship $R_p = 1.037\ S_2^{0.023}$. (*b*) For $g = 3$, we separated those phenotypes belonging to $\mathcal{P}_2$ (green circles, above) from the rest (blue circles, below). Both sets show a power law relationship between phenotypic robustness and phenotypic abundance: $R_p = 1.790\ S_3^{0.013}$ for phenotypes in $\mathcal{P}_2$ (green line) and $R_p = 0.805\ S_3^{0.023}$ for the remaining 25 717 phenotypes (straight lines). Grey area encompasses two standard deviations, and the fits were obtained using the least-squares method. (Online version in colour.)

fragmentation of neutral networks is due to the HP model that underlies protein folding in toyLIFE: there are no neutral mutations within proteins (electronic supplementary material, §S1 and figure S3). Most mutations in a protein sequence yield proteins that do not fold, and few mutations yield one functional protein from another. Remember that each phenotype is defined by a list of metabolites that a genotype is able to catabolize. A given phenotype can be encoded by more than one protein combination, which will be characterized by different interactions and regulatory functions, but will catabolize the same metabolites. When a phenotype is generated by several protein combinations, it will be difficult to mutate from one combination to another, and as a consequence these combinations will usually be found in disjointed connected components. In fact, the number of connected components associated with a phenotype is positively correlated with the number of protein combinations that generate it (electronic supplementary material, §S7).

For $g \geq 2$, we can estimate the distribution of neutral networks in genotype space using neutral random walks: starting at a randomly chosen genotype, we perform a mutation on it. If the resulting mutant genotype belongs to the same neutral network—if it encodes the same phenotype—the mutation is accepted. The random walk continues when we mutate the new genotype again. If the mutant genotype does not belong to the neutral network, the mutation is rejected, and we try to find a new neutral neighbour for the original genotype (this process will not work if the starting genotype does not have neutral neighbours, a rare case). We performed 1000 neutral random walks of length 10 000 for genotype sizes $g = 2$ to $g = 5$ (electronic supplementary material, figure S15). At each time step t, we computed $d_H(g_0, g_t)$, the Hamming distance (normalized number of different positions) between the original genotype g_0 and the genotype visited at time t, g_t. $d_H(g_0, g_t)$ is a random variable for each t, and so we can compute its average and standard deviation (electronic supplementary material, figure S15). If there were no restrictions to the nodes that can be visited in a random walk, we would expect $d_H(g_0, g_t) \to 0.5$ when $t \to \infty$. In other words, if there are no restrictions, the correlation between g_0 and g_t is lost when t grows, and the distance between them tends to the value it would have, on average, if we randomly picked two genotypes from the network. Thus, the evolution of $d_H(g_0, g_t)$ is a good measure of the size and extension of neutral networks in genotype space. For $g = 2$, $\langle d_H(g_0, g_t)\rangle \sim 0.25$ when $t \to \infty$, implying that networks do not extend very far. Considering that the total genotype space has diameter 40, this means that the average distance between the initial genotype and the final one is close to 10. This is not a very high value, and it is consistent with our previous analysis showing that neutral networks in $g = 2$ tend to be fragmented and small. For $g > 2$, $\langle d_H(g_0, g_t)\rangle \sim 0.4$ when $t \to \infty$, which implies that the fragmented networks of $g = 2$ space are becoming more connected as g grows, facilitating the navigability of genotype space. This suggests that neutral networks for $g > 2$ span large fractions of genotype space, a result consistent with other genotype–phenotype maps.

A different way to estimate the diameter of a neutral network is to perform neutral random walks in which we force $d_H(g_t, g_{t+1}) > d_H(g_{t-1}, g_t)$. That is, in addition to imposing that a mutation is neutral in order to accept it, we also require it to increase the distance to the original genotype. More specifically, the process is computed as follows. We randomly choose a genotype, and perform mutations on it, increasing the distance every time step, until this distance can increase no longer—if, after a large number of trials, we cannot find a neutral mutant that is farther apart from the original genotype, we stop the process. We will denote the final distance obtained in such random walks by d_∞. For $g = 2$ and $g = 3$ we randomly sampled 10 000 genotypes, whereas for $g = 4$ and $g = 5$ we sampled 1000 genotypes (figure 2*d*; electronic supplementary material, figure S16). Consistent with previous results, random walks did not get very far for $g = 2$, reaching an average final distance $\langle d_\infty\rangle \sim 0.28$. For $g > 2$, the final distance d_∞ increases. This result confirms the previous observation that navigability in these genotype spaces is enhanced. For $g = 3$, $\langle d_\infty\rangle$ is a little over 0.5, while for $g = 4$ and $g = 5$ it reaches 0.6 and 0.7, respectively. In fact, the growth of $\langle d_\infty\rangle$ with g is very similar to the growth of $\langle R\rangle$ obtained in figure 2*c*. In that case, we had $\langle R\rangle = 0.895 - 1.392/g$. Here, it is $\langle d_\infty\rangle = 0.965 - 1.354/g$ (figure 2*d*). Unsurprisingly, the similarity of the fits implies a linear relationship between $\langle d_\infty\rangle$ and $\langle R\rangle$: $\langle d_\infty\rangle = 0.094 + 0.972\langle R\rangle$ (figure 2*e*), very close to the identity function. This result has several implications. First, as g grows, neutral networks are more and more connected, and they span larger fractions of genotype space. It is easier to get from one extreme of the genotype network to the other without changing the phenotype. Secondly, this increased connectivity is due to the increase in robustness:

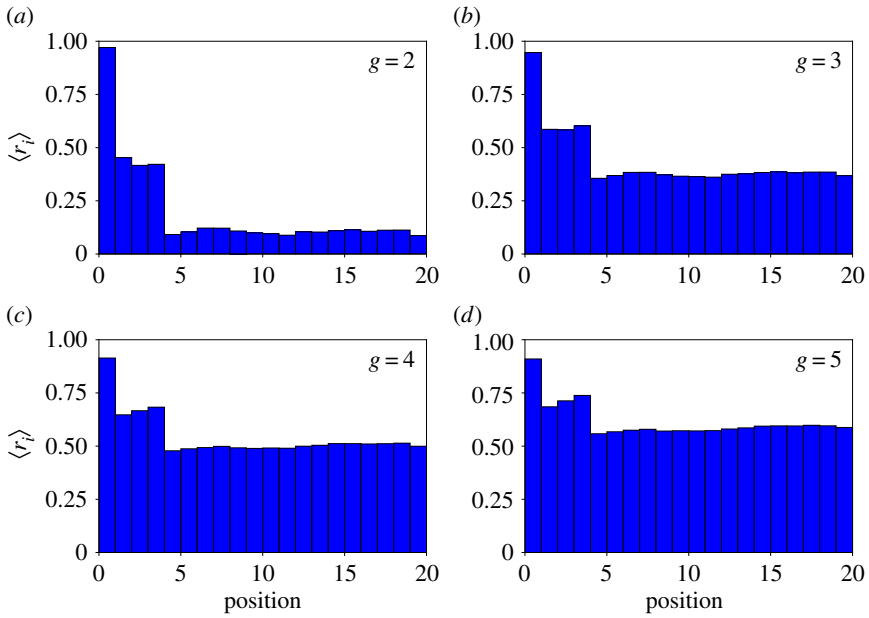


Figure 4. Different positions in the genome have different neutralities. We sampled 10^7 genotypes for $g = 2$ (*a*) and $g = 3$ (*b*) and 10^3 genotypes for $g = 4$ (*c*) and $g = 5$ (*d*), and measured r_i for $i = 1, \ldots, 20$. For each i we then computed $\langle r_i \rangle$ and plotted them versus genomic position. Note the high robustness of the first position in the promoter region, and the low robustness in the coding regions. (Online version in colour.)

the robustness of a genotype is a good predictor for the size of the connected component it belongs to. This can be easily explained in light of our previous discussion on robustness. Adding a new gene to a genotype will endow the latter with an average of 18 new neutral mutations with which to explore genotype space (figure 2*b*). Because the new gene will not interfere with the phenotype with a high probability, it follows that we can mutate most of its nucleotides, one by one, getting farther away from the original genotype. In other words, new genes in toyLIFE allow for increased navigability of genotype space, because they are mostly junk genes. As we will see later on, this property will have important consequences for evolvability.

The fact that robustness is a good predictor for the size of a genotype's connected component can be combined with the positive correlation between the logarithmic abundance of a phenotype and the size of its largest connected component (electronic supplementary material, §S7) to deduce the linear relationship between the logarithm of phenotype abundance and phenotypic robustness, giving yet another heuristic argument for this relationship. We now turn to compute it explicitly.

Phenotypic robustness is defined as the average of genotypic robustness for all genotypes encoding a phenotype $\mathcal{P}_i$, that is

$$R_{\mathcal{P}_i} = \frac{1}{|\mathcal{P}_i|} \sum_{g \in \mathcal{P}_i} R_g,$$

where $|\mathcal{P}_i|$ is the number of genotypes encoding $\mathcal{P}_i$. For $g = 2$ and $g = 3$ we sampled 10^7 genotypes and computed their robustness. We then assigned each genotype to its corresponding phenotype and averaged the values of robustness for all genotypes encoding each phenotype. Note that this procedure samples abundant phenotypes more often. For $g = 2$, we find a good fit to a linear relationship between the logarithm of phenotype abundance and estimated phenotypic robustness (figure 3*a*). For $g = 3$, we identified those phenotypes belonging to $\mathcal{P}_2$ (being the most abundant, they were sampled the most) in green, and the rest in blue (figure 3*b*). The figure shows separate relationships between the logarithm of phenotype abundance and phenotypic robustness. The two sets of phenotypes cluster in two different groups, confirming once more the idea that these two sets are qualitatively different. Phenotypes belonging to $\mathcal{P}_2$ are much more robust, as a result of them having one spare junk gene.

4. Robustness and position in genotype

Instead of considering the degree of a node in a neutral network, we can focus on the neutrality of a given position of the genotype. A genotype formed by g genes can be thought of as a binary string of length $20g$—remember that genes in toyLIFE have 20 nucleotides, the first four forming the promoter region and the remaining 16 constituting the coding region. For a given sequence, the position $i = 1, \ldots, 20g$ can either be neutral or not—that is, when we mutate that position, we can get a new genotype with the same phenotype or not. We can thus define the random variable

$$r_i = \begin{cases} 1 & \text{if } i \text{ is a neutral position,} \\ 0 & \text{otherwise.} \end{cases}$$

Because r_i is a random variable, we can sample it and estimate its mean. Differences between positions may yield insights into the details of the genotype–phenotype correspondence in toyLIFE—i.e. some positions may always be neutral, or always constrained. This is what we have done in figure 4, for genotype sizes $g = 2$ to $g = 5$. We sampled 10^7 genotypes for $g = 2$ and $g = 3$, and 10^3 genotypes for $g = 4$ and $g = 5$, and computed r_i for every $i = 1, \ldots, 20g$ and every genotype. The order of genes does not matter in toyLIFE by construction—implying $\langle r_i \rangle = \langle r_{i+20h} \rangle$, for any $h = 0, 1, \ldots, g - 1$—so we are interested in the values of robustness for each gene. This is why in figure 4 we only show the average values $\langle r_i \rangle$ for $0 \leq i < 20$. Note that promoter regions tend to be more robust than coding regions. This is partly due to the lack of robustness in the version of the HP model that underlies

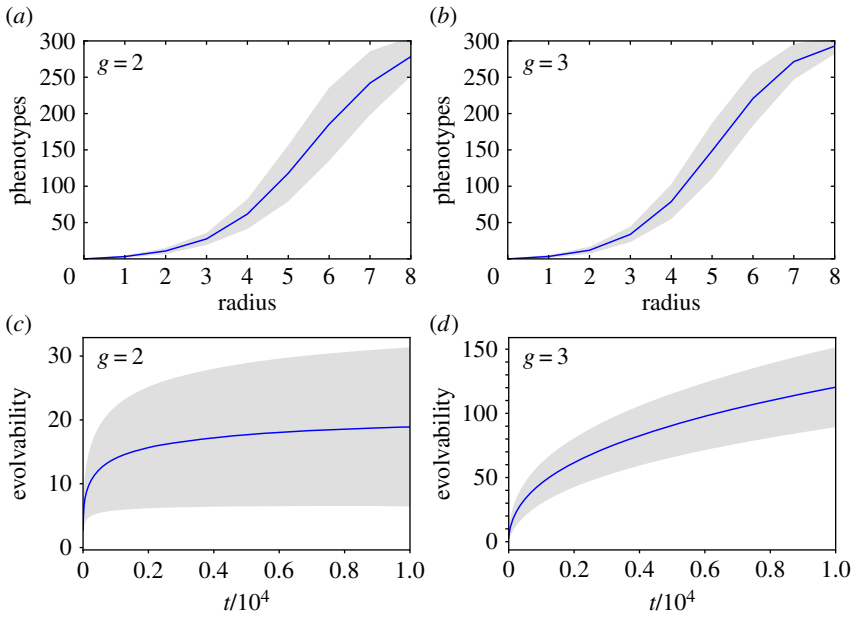


Figure 5. Evolvability in toyLIFE. A genotype–phenotype map has the shape space covering property if, given a phenotype, we only need to explore a small radius around a sequence encoding that phenotype in order to find the most common phenotypes. We tested this property in toyLIFE by sampling 100 genotypes for $g = 2$ (*a*) and $g = 3$ (*b*), and counting how many of the 300 most common phenotypes appeared in a radius of distance 8 around a given genotype. The results are consistent with shape space covering. For $g = 2$ (*c*) and $g = 3$ (*d*), we also measured the cumulative number of phenotypes in the neighbourhood of a neutral random walk—evolvability—for 10 000 genotypes. Evolvability is much higher for $g = 3$. Lines show average values, while grey areas encompass two standard deviations. (Online version in colour.)

protein folding in toyLIFE (electronic supplementary material, §S1 and figure S3). However, note that the superposition of regulatory and metabolic levels of the phenotype makes the average robustness of coding regions grow, in spite of the non-robust protein folding model. For $g = 4$ and $g = 5$, the average robustness in these regions reaches values as high as 0.5 (remember that these genotypes tend to have junk genes that increase overall robustness as g grows).

Inside the promoter region, which only affects gene regulation, the first position is particularly robust. This means that the regulatory changes it induces have no phenotypic effect at the metabolic level. This may be due to two reasons: either changes in the first position of the promoter region do not affect the regulatory function—the temporal pattern of gene expression determined by the interactions among proteins—or changes in the regulatory function rarely alter the metabolic phenotype. We performed the following simple test of these hypotheses. For each position in the promoter region, we sampled 10^4 genotypes of size $g = 3$. We then mutated that position and computed the new regulatory function and the new metabolic phenotype. From all 10^4 mutations in the first position, 40% were neutral in both the regulatory and the metabolic sense, 54% affected the regulatory function but did not affect the metabolic phenotype, and the remaining 6% changed both—this means that the robustness for the first position in this sample was 94%. For the rest of the positions, 27% of the mutations did not alter either the regulatory function or the metabolic phenotype, 32% changed regulation but not metabolism and 41% changed both. This corresponds to a robustness of 59%, consistent with what we observed in figure 4. In other words, for the first position only 9% of the mutations that affected regulation had any effect on the metabolic phenotype. In addition, 40% of mutations did not affect the regulatory function at all. For the rest of the positions, however, the number of mutations that altered regulation, 73%, was higher. Among these, roughly 55% had an effect on phenotype as well. So both reasons posited above are at work: not only is the number of mutations affecting regulatory function lower in the first position of the promoter region, but also when these mutations do alter the regulatory function, they rarely change the phenotype.

The lower robustness of coding regions, compared to promoter regions, is correlated with a higher evolvability, as will be discussed in the next Section.

5. Accessibility and evolvability

So far, we have limited our discussion of the properties of the genotype–phenotype map in toyLIFE to the abundance of phenotypes and the organization of their neutral networks, without paying any attention to the connections between different phenotypes. In this section, we will focus on the latter question, which amounts to studying evolvability, or how accessible phenotypes are.

The neutral networks of different phenotypes tend to be highly interwoven in most computational genotype–phenotype maps, so that connections between them are very common. The Vienna RNA group described a property of RNA neutral networks called shape space covering [7,9]. It implies that one can find most common phenotypes a few mutations away from any given genotype. We checked for the existence of this property in toyLIFE. We sampled 100 genotypes for $g = 2$ and $g = 3$ and computed the phenotypes of all neighbours at distances 1 to 8. We observed how many of the 300 most common phenotypes appeared in this set of neighbours. The results are shown in figure 5*a*,*b*. For both $g = 2$ and $g = 3$ and most sampled genotypes, the number of phenotypes discovered after eight mutations was close to 300. This implies that toyLIFE also shares the shape space covering

property: most phenotypes are just a few mutations away from any given phenotype. Observe, however, that for $g = 2$ this means a higher relative distance compared with $g = 3$—remember that the diameter of this network is 40—and that the number of phenotypes discovered at that distance is lower by comparison.

Shape space covering means that phenotypes are easily accessible from each other through a few number of mutations. A relevant detail in the metabolic genotype–phenotype map in toyLIFE is that this accessibility is due only to mutations in proteins. If we mutate only the promoters, the number of visited phenotypes is never larger than 2, regardless of the distance. For $g = 2$, we can give a clear explanation to this peculiarity: of the 135 318 pairs of proteins that yield a metabolic function, only 16 yield two different metabolic phenotypes when combined with different promoters. The rest are able to generate only one phenotype. Changing only the promoters will not affect the metabolic function, and will not help in finding new phenotypes. This is consistent with the high robustness of the promoter region.

Another way to study evolvability is to compute the connections between different phenotypes directly. We say that two phenotypes are connected if at least two genotypes from each phenotype are one point mutation away from each other. We can then create a network of phenotypes, whose nodes will be the phenotypes themselves, and the edges the connections between them. This network of phenotypes is undirected and weighted—the weight of an edge between two phenotypes is the sum of all point mutations connecting two genotypes encoding each phenotype. This network admits self-loops, whose weight is twice the number of edges connecting genotypes encoding a single phenotype—in other words, it is the sum of the degrees of all the nodes encoding that phenotype, which is proportional to the phenotype's robustness. For $g = 2$, where we can compute the whole network of genotypes with their corresponding phenotypes, we can build this phenotype network exhaustively. The network is formed by a giant component that includes 767 phenotypes. We also find six additional tiny components, five of them with just one phenotype and the remaining one with three phenotypes (electronic supplementary material, §S9). Thus, for $g = 2$, some phenotypes will be unreachable by point mutations from other phenotypes. For $g = 3$, we cannot build the phenotype network exhaustively, but resorting to a numerical approximation using random walks (electronic supplementary material, §S9), we can estimate the network of connections between the 775 phenotypes in $\mathcal{P}_2$—in order to study how the addition of one gene alters the connections between these phenotypes. The results show that all phenotypes in $\mathcal{P}_2$ now belong to one giant component (electronic supplementary material, §S9). The number of connections between phenotypes has greatly increased as well. This is again due to the additional junk genes. They do not only increase robustness, but also allow for increased connections between phenotypes.

This increased connectivity can also be measured in an alternative way. In previous work [5], evolvability has been estimated as the number of new phenotypes discovered in a neutral random walk along a neutral network. In figure 5*c*,*d*, we have performed such an analysis for 10 000 genotypes for $g = 2$ and $g = 3$. The results show that evolvability is much higher for $g = 3$. While the number of discovered phenotypes almost stops growing for $g = 2$, it grows quickly in $g = 3$, and to a much higher value than for $g = 2$. Again, this is due to the higher average number of connections between phenotypes for $g = 3$ (electronic supplementary material, §S9).

6. Discussion and conclusion

Throughout this article, we have explored the properties of the metabolic genotype–phenotype map in toyLIFE. This map is highly degenerate, with many more genotypes than phenotypes, and large neutral networks traversing genotype space. The distribution of phenotype abundances is very heterogeneous, and more abundant phenotypes tend to be more robust. Common phenotypes are easily accessed from each other, and large neutral networks allow for a fast exploration of phenotype space.

All of these properties have been described in other genotype–phenotype maps [7–31,33,34]. This is somewhat striking, given that the genotype–phenotype map in toyLIFE is more complex than the rest of these models. It is the only model that incorporates intermediate levels of phenotypic expression: genes are first translated into proteins, which fold and interact with each other, generating complex regulatory networks that will determine the metabolic capacities of a genotype. And yet the main properties shared by the rest of genotype–phenotype maps appear here as well.

Two particular results stand out. The first is the log-normal distribution of phenotype abundances, which has also been observed in RNA [14] and predicted for simple combinatorial genotype–phenotype maps [34]. The second is the positive correlation between phenotypic robustness and the logarithm of phenotype abundance, which has also been described before [13,30,31,34]. The fact that these two relationships (as well as other phenomena, such as shape space covering) are so widespread points to a general property of these maps, which must be related to combinatorics and network theory. Previous work [34] has shown that, when the abundance of a phenotype can be inferred from the genotype sequence in simple genotype–phenotype maps, we can use combinatorial arguments to explain the appearance of a log-normal distribution and the linear relationship between phenotypic robustness and the logarithm of phenotype abundance. These arguments would explain the presence of these properties in the case of RNA, but do not seem to be easily translatable to toyLIFE. We need to devote more efforts into understanding this seemingly general property of genotype–phenotype maps.

The high robustness and evolvability of the metabolic phenotype in toyLIFE, particularly when genome size increases, is remarkable because our model is built on a particularly non-robust, non-evolvable version of the HP protein-folding model. Proteins in toyLIFE are sequences of 16 amino-acids that fold on a 4×4 lattice. One protein is defined by its perimeter and its folding energy: this is a very different definition of protein than the one used in most versions of the HP model [15–20], which define a protein by its folded structure. As we have already mentioned, there are no neutral networks in the protein space in toyLIFE, and evolvability is very limited. However, when these proteins are paired to interact with each other, generating regulatory networks and performing metabolic functions, the resulting genotypes are robust and evolvable. So it would seem that adding levels of expression to a phenotype enhances both robustness and evolvability. This hypothesis could be tested, for instance, by adding another level of complexity to the phenotype in toyLIFE and checking how both of these

properties are changed. If there is indeed a relationship between complexity and both robustness and evolvability, this result would suggest that more complex genotype–phenotype maps could have an evolutionary advantage.

Alternatively, we could change the folding process in toyLIFE and allow for promiscuous proteins—proteins that can fold in different shapes with the same energy [42]. This would increase connectivity in protein space and would affect the levels of robustness and evolvability at the metabolic level. Determining the extent of this change would give us some insight into the relationship between different levels of phenotypic expression.

On a related note, our results show that adding genes to toyLIFE genotypes increases both robustness and evolvability. Neutral networks of two-gene genotypes are not very navigable, reaching only a small portion of genotype space, and connecting with a small number of adjacent phenotypes. However, adding a new gene to the genotype changes everything: now phenotypes are easily accessed from each other and neutral networks span genotype space. Robustness, as we haven seen, keeps growing with genome size. The explanation behind this fact is that most of the new genes will not alter the metabolic phenotype, and will act as junk genes. However, they can mutate without restriction, enhancing the navigability of a neutral network. Increased navigability allows for increased connectivity between phenotypes, thus enhancing evolvability [43]. In other words, junk genes have creative potential, in the sense that they allow populations to explore a given neutral network, and then encounter new, unexplored phenotypes. This is interesting because it extends the usefulness of redundancy in complex genomes [44,45] to include seemingly inert elements, whose only function is to increase robustness and evolvability. It is also reminiscent of the abundance of introns and non-coding DNA in eukaryotic genomes [46]: if this non-functional DNA also enhances robustness and evolvability in living cells, this would suggest new arguments for the maintenance of junk DNA.

Finally, the appearance of junk genes is possibly a result of the fact that interactions in toyLIFE are limited to be pairwise. There are neither trimers nor tetramers in toyLIFE, only dimeric proteins. Only one protein or dimer can interact with a metabolite at a given moment, and so on. As a consequence, when a new gene is added to a two-gene genotype that performs a metabolic function, it will have little potential to create new functions. In fact, one would expect the opposite: that adding a new gene would disrupt the existing interactions, thus yielding a non-viable metabolic phenotype. However, this is not what we observe. When adding new genes to two-gene genotypes, most genotypes keep their original function. We will need to perform a more detailed exploration of this phenomenon in order to clarify the reasons behind it.

As a final comment, note that we have not defined fitness for toyLIFE. Previous work [47] suggests that the fitness landscape appearing from complex genotype–phenotype maps is highly rugged and constrains evolutionary paths. Further work with toyLIFE should explore possible definitions of fitness and their corresponding fitness landscapes.

Data accessibility. There are no data associated with this paper.

Authors' contributions. P.C., S.M. and J.A.C. conceived the study. P.C. performed research and data analysis. P.C., A.W., S.M. and J.A.C. discussed the results, wrote the paper and approved the final version of the manuscript.

Competing interests. We declare we have no competing interests.

Funding. P.C., S.M. and J.A.C. acknowledge support by the Spanish Ministerio de Economía y Competitividad and FEDER funds of the EU through grants VARIANCE (FIS2015-64349-P) (P.C. and J.A.C.) and ViralESS (FIS2014-57686-P) (S.M.). A.W. acknowledges support by ERC Advanced Grant 739874, by Swiss National Science Foundation grant 31003A_146137, by an EpiphysX RTD grant from SystemsX.ch, as well as by the University Priority Research Program in Evolutionary Biology at the University of Zurich. P.C. acknowledges additional support from EMBO grant ASTF 652-2014.

Acknowledgments. We acknowledge the revision of six anonymous reviewers, which has helped in improving this paper.

References

1. Alberch P. 1991 From genes to phenotype: dynamical systems and evolvability. *Genetica* **84**, 5–11. (doi:10.1007/BF00123979)
2. Eldredge N, Gould SJ. 1972 Punctuated equilibria: an alternative to phyletic gradualism. In Models in Paleobiology (ed. TJM Schopf), pp. 82–115. San Francisco, CA: Freeman Cooper.
3. Gould SJ, Eldredge N. 1977 Punctuated equilibria: the tempo and mode of evolution reconsidered. *Paleobiology* **3**, 115–151. (doi:10.1017/S0094837300005224)
4. Maynard Smith J, Burian R, Kauffman S, Alberch P, Campbell J, Goodwin B, Lande R, Raup D, Wolpert L. 1985 Developmental constraints and evolution: a perspective from the Mountain Lake conference on development and evolution. *Q. Rev. Biol.* **60**, 265–287. (doi:10.1086/414425)
5. Wagner A. 2011 *The origins of evolutionary innovations*. Oxford, UK: Oxford University Press.
6. Nei M. 2013 *Mutation-driven evolution*. Oxford, UK: Oxford University Press.
7. Schuster P, Fontana W, Stadler PF, Hofacker IL. 1994 From sequences to shapes and back: a case study in RNA secondary structures. *Proc. R. Soc. Lond. B* **255**, 279–284. (doi:10.1098/rspb.1994.0040)
8. Grüner W, Giegerich R, Strothmann D, Reidys C, Weber J, Hofacker IL, Stadler PF, Schuster P. 1996 Analysis of RNA sequence structure maps by exhaustive enumeration. I. Neutral networks. *Monatsh. Chem.* **127**, 355–374. (doi:10.1007/BF00810881)
9. Grüner W, Giegerich R, Strothmann D, Reidys C, Weber J, Hofacker IL, Stadler PF, Schuster P. 1996 Analysis of RNA sequence structure maps by exhaustive enumeration II. Structures of neutral networks and shape space covering. *Monatsh. Chem.* **127**, 375–389. (doi:10.1007/BF00810882)
10. Huynen MA. 1996 Exploring phenotype space through neutral evolution. *J. Mol. Evol* **43**, 165–169. (doi:10.1007/BF02338823)
11. Fontana W, Schuster P. 1998 Continuity in evolution: on the nature of transitions. *Science* **280**, 1451–1455. (doi:10.1126/science.280.5368.1451)
12. Jörg T, Martin OC, Wagner A. 2008 Neutral network sizes of biological RNA molecules can be computed and are not atypically small. *BMC Bioinformatics* **9**, 1. (doi:10.1186/1471-2105-9-1)
13. Aguirre J, Buldú JM, Stich M, Manrubia SC. 2011 Topological structure of the space of phenotypes: the case of RNA neutral networks. *PLoS ONE* **6**, e26324. (doi:10.1371/journal.pone.0026324)
14. Dingle K, Schaper S, Louis AA. 2015 The structure of the genotype–phenotype map strongly constrains the evolution of non-coding RNA. *Interface Focus* **5**, 20150053. (doi:10.1098/rsfs.2015.0053)
15. Lau KF, Dill KA. 1989 A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules* **22**, 3986–3997. (doi:10.1021/ma00200a030)
16. Lipman DJ, Wilbur WJ. 1991 Modelling neutral and selective evolution of protein folding. *Proc. R. Soc. Lond. B* **245**, 7–11. (doi:10.1098/rspb.1991.0081)
17. Li H, Helling R, Tang C, Wingreen N. 1996 Emergence of preferred structures in a simple model

of protein folding. *Science* **273**, 666–669. (doi:10.1126/science.273.5275.666)

18. Bornberg-Bauer E. 1997 How are model protein structures distributed in sequence space? *Biophys. J.* **73**, 2393–2403. (doi:10.1016/S0006-3495(97)78268-7)
19. Bastolla U, Roman HE, Vendruscolo M. 1999 Neutral evolution of model proteins: diffusion in sequence space and overdispersion. *J. Theor. Biol.* **200**, 49–64. (doi:10.1006/jtbi.1999.0975)
20. Irbäck A, Troein C. 2002 Enumerating designing sequences in the HP model. *J. Biol. Phys.* **28**, 1–15. (doi:10.1023/A:1016225010659)
21. Ciliberti S, Martin OC, Wagner A. 2007 Innovation and robustness in complex regulatory gene networks. *Proc. Natl Acad. Sci. USA* **104**, 13 591–13 596. (doi:10.1073/pnas.0705396104)
22. Cotterell J, Sharpe J. 2010 An atlas of gene regulatory networks reveals multiple three-gene mechanisms for interpreting morphogen gradients. *Mol. Sys. Biol.* **6**, 425. (doi:10.1038/msb.2010.74)
23. Payne JL, Moore JH, Wagner A. 2014 Robustness, evolvability, and the logic of genetic regulation. *Artif. Life* **20**, 111–126. (doi:10.1162/ARTL_a_00099)
24. Jiménez A, Cotterell J, Munteanu A, Sharpe J. 2015 Dynamics of gene circuits shapes evolvability. *Proc. Natl Acad. Sci. USA* **112**, 2103–2108. (doi:10.1073/pnas.1411065112)
25. Rodrigues JFM, Wagner A. 2009 Evolutionary plasticity and innovations in complex metabolic reaction networks. *PLoS Comp. Biol.* **5**, e1000613. (doi:10.1371/journal.pcbi.1000613)
26. Rodrigues JFM, Wagner A. 2011 Genotype networks, innovation, and robustness in sulfur metabolism. *BMC Sys. Biol.* **5**, 39. (doi:10.1186/1752-0509-5-39)
27. Wagner A, Andriasyan V, Barve A. 2014 The organization of metabolic genotype space facilitates adaptive evolution in nitrogen metabolism. *J. Mol. Biochem.* **3**, 2–13.
28. Hosseini SR, Barve A, Wagner A. 2015 Exhaustive analysis of a genotype space comprising 10 15 central carbon metabolisms reveals an organization conducive to metabolic innovation. *PLoS Comput. Biol.* **11**, e1004329. (doi:10.1371/journal.pcbi.1004329)
29. Johnston IG, Ahnert SE, Doye JP, Louis AA. 2011 Evolutionary dynamics in a simple model of self-assembly. *Phys. Rev. E* **83**, 066105. (doi:10.1103/PhysRevE.83.066105)
30. Greenbury SF, Johnston IG, Louis AA, Ahnert SE. 2014 A tractable genotype-phenotype map modelling the self-assembly of protein quaternary structure. *J. R. Soc. Interface* **6**, 20140249. (doi:10.1098/rsif.2014.0249)
31. Greenbury SF, Schaper S, Ahnert SE, Louis AA. 2016 Genetic correlations greatly increase mutational robustness and can both reduce and enhance evolvability. *PLoS Comp. Biol.* **12**, e1004773. (doi:10.1371/journal.pcbi.1004773)
32. Arias CF, Catalán P, Manrubia S, Cuesta JA. 2014 toyLIFE: a computational framework to study the multi-level organisation of the genotype-phenotype map. *Sci. Rep.* **4**, 7549. (doi:10.1038/srep07549)
33. Greenbury S, Ahnert S. 2015 The organization of biological sequences into constrained and unconstrained parts determines fundamental properties of genotype–phenotype maps. *J. R. Soc. Interface* **12**, 20150724. (doi:10.1098/rsif.2015.0724)
34. Manrubia S, Cuesta JA. 2017 Distribution of genotype network sizes in sequence-to-structure genotype–phenotype maps. *J. R. Soc. Interface* **14**, 20160976. (doi:10.1098/rsif.2016.0976)
35. Ebner M, Shackleton M, Shipman R. 2001 How neutral networks influence evolvability. *Complexity* **7**, 19–33. (doi:10.1002/cplx.10021)
36. Fortuna MA, Zaman L, Ofria C, Wagner A. 2017 The genotype-phenotype map of an evolving digital organism. *PLoS Comp. Biol.* **13**, e1005414. (doi:10.1371/journal.pcbi.1005414)
37. Schaper S, Louis AA. 2014 The arrival of the frequent: how bias in genotype-phenotype maps can steer populations to local optima. *PLoS ONE* **9**, e86635. (doi:10.1371/journal.pone.0086635)
38. Maynard Smith J. 1970 Natural selection and the concept of a protein space. *Nature* **225**, 563–564. (doi:10.1038/225563a0)
39. Ciliberti S, Martin OC, Wagner A. 2007 Robustness can evolve gradually in complex regulatory gene networks with varying topology. *PLoS Comp. Biol.* **3**, e15. (doi:10.1371/journal.pcbi.0030015)
40. Dall'Olio GM, Bertranpetit J, Wagner A, Laayouni H. 2014 Human genome variation and the concept of genotype networks. *PLoS ONE* **9**, e99424. (doi:10.1371/journal.pone.0099424)
41. Barve A, Wagner A. 2013 A latent capacity for evolutionary innovation through exaptation in metabolic systems. *Nature* **500**, 203–206. (doi:10.1038/nature12301)
42. Khersonsky O, Tawfik DS. 2010 Enzyme promiscuity: a mechanistic and evolutionary perspective. *Ann. Rev. Biochem.* **79**, 471–505. (doi:10.1146/annurev-biochem-030409-143718)
43. Wagner A. 2008 Robustness and evolvability: a paradox resolved. *Proc. R. Soc. B* **275**, 91–100. (doi:10.1098/rspb.2007.1137)
44. Daniels BC, Chen YJ, Sethna JP, Gutenkunst RN, Myers CR. 2008 Sloppiness, robustness, and evolvability in systems biology. *Curr. Opin. Biotech.* **19**, 389–395. (doi:10.1016/j.copbio.2008.06.008)
45. Whitacre J, Bender A. 2010 Degeneracy: a design principle for achieving robustness and evolvability. *J. Theor. Biol.* **263**, 143–153. (doi:10.1016/j.jtbi.2009.11.008)
46. Doolittle WF. 2013 Is junk DNA bunk? A critique of ENCODE. *Proc. Natl Acad. Sci. USA* **110**, 5294–5300. (doi:10.1073/pnas.1221376110)
47. Khatri BS, McLeish TC, Sear RP. 2009 Statistical mechanics of convergent evolution in spatial patterning. *Proc. Natl Acad. Sci. USA* **106**, 9564–9569. (doi:10.1073/pnas.0812260106)

Supplementary Material

Adding levels of complexity enhances robustness and evolvability in a multi-level genotype-phenotype map

Pablo Catalán, Andreas Wagner, Susanna Manrubia and José A. Cuesta

Journal of the Royal Society Interface

Contents

1 toyLIFE

toyLIFE was originally presented in [1]. We give here its main details, with slight modifications in the definition of the model.

1.1 Building blocks: genes, proteins, metabolites

The basic building blocks of toyLIFE are toyNucleotides (toyN), toyAminoacids (toyA), and toySugars (toyS). Each block comes in two flavors: hydrophobic (H) or polar (P). Random polymers of basic blocks constitute toyGenes (formed by 20 toyN units), toyProteins (chains of 16 toyA units), and toyMetabolites (sequences of toyS units of arbitrary length). These elements of toyLIFE are defined on two-dimensional space (Supplementary Figure 1).

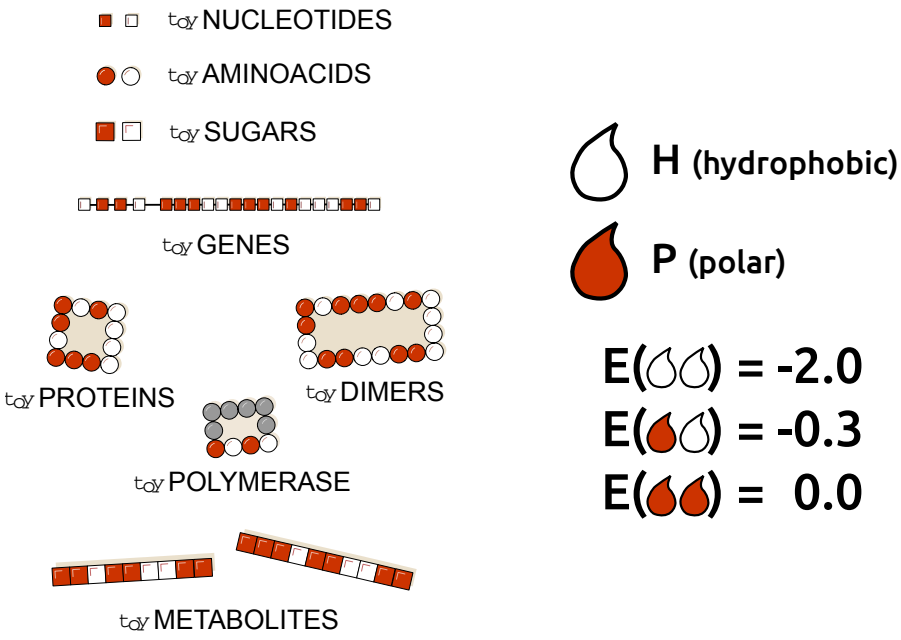
toyGenes

toyGenes are composed of a 4-toyN promoter region followed by a 16-toyN coding region. There are 2^4 different promoters and 2^{16} coding regions, leading to $2^{20} \approx 10^6$ toyGenes. An ensemble of toyGenes forms a genotype. If the toyGene is expressed, it will produce a chain of 16 toyA that represents a toyProtein. Translation follows a straightforward rule: H (P) toyN translate into H (P) toyA. Point mutations in toyLIFE are easy to implement: they are changes in one of the nucleotides in one of the genes in the genotype. If the sequence has a H toyN in that position, then a mutation will change it to a P toyN, and vice versa.

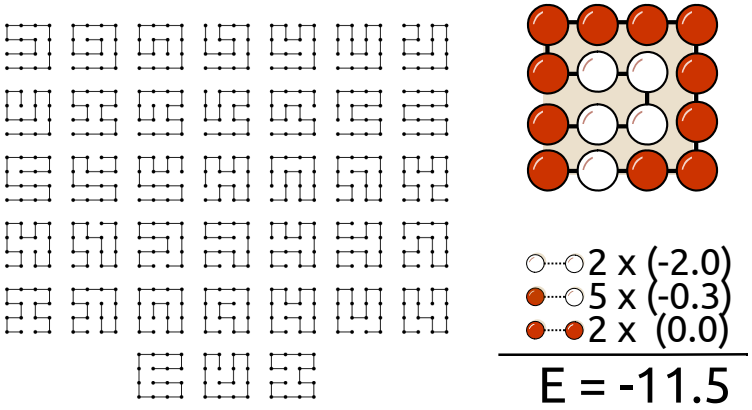
toyProteins

toyProteins correspond to the minimum energy, maximally compact folded structure of the 16 toyA chain arising from a translated toyGene. Their folded configuration is calculated through the hydrophobic-polar (HP) protein lattice model [2, 3].

We only consider maximally compact structures. That is, every toyProtein must fold on a 4×4 lattice, following a self-avoiding walk (SAW) on it. After accounting for symmetries —rotations and reflections—, there are only 38



Supplementary Figure 1: Building blocks and interactions defining toyLIFE. The three basic building blocks of toyLIFE are toyNucleotides, toyAminoacids, and toySugars. They can be hydrophobic (H, white) or polar (P, red), and their random polymers constitute toyGenes, toyProteins, and toyMetabolites. The toyPolymerase is a special polymer that will have specific regulatory functions. These polymers will interact between each other following an extension of the HP model (see text), for which we have chosen the interaction energies $E_{\mathrm{HH}} = -2$, $E_{\mathrm{HP}} = -0.3$ and $E_{\mathrm{PP}} = 0$ [2].



Supplementary Figure 2: Protein folding in ToyLIFE. toyProteins fold on a 4×4 lattice, following a self-avoiding walk (SAW). Discarding for symmetries, there are 38 SAWs (left). For each binary sequence of length 16, we fold it into every SAW and compute its folding energy, following the HP model. For instance, we fold the sequence PHPPPPPPPPPHHHHP into one of the SAWs and compute its folding energy (right). There are two HH contacts, five HP contacts and two PP contacts —we only take into account contacts between non-adjacent toyAminoacids. Summing all this contacts with their corresponding energies, we obtain a folding energy of -11.5. Repeating this process for every SAW, we obtain the minimum free structure.

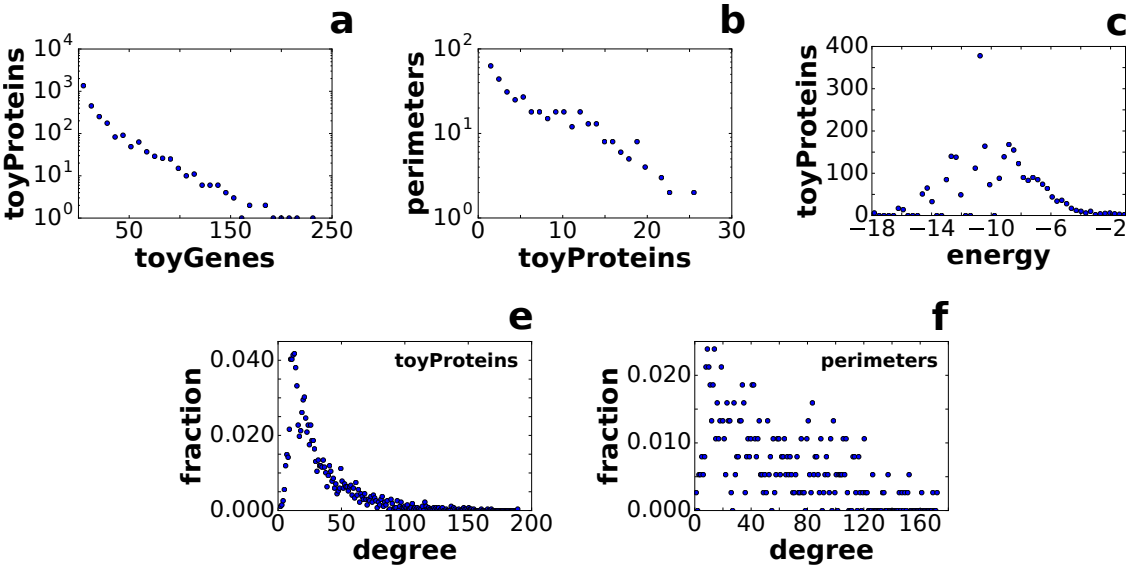
SAWs on that lattice (Supplementary Figure 2).

The energy of a fold is the sum of all pairwise interaction energies between toyA that are not contiguous along the sequence. Pairwise interaction energies are $E_{\mathrm{HH}} = -2$, $E_{\mathrm{HP}} = -0.3$ and $E_{\mathrm{PP}} = 0$, following the conditions set in [2] that $E_{\mathrm{PP}} > E_{\mathrm{HP}} > E_{\mathrm{HH}}$ (Supplementary Figure 2). toyProteins are identified by their folding energy and their perimeter. If there is more than one fold with the same minimum energy, we select the one with fewer H toyAminoacids in the perimeter. If still there is more than one fold fulfilling both conditions, we discard that protein by assuming that it is intrinsically disordered and thus non-functional. Note, however, that sometimes different folds yield the same folding energy and the same perimeter. In those cases, we do not discard the resulting toyProtein [1]. Out of $2^{16} = 65,536$ possible toyProteins, $12,987$ do not yield unique folds. We find $2,710$ different toyProteins with 379 different perimeters. Not all toyProteins are equally abundant: although every toyProtein is coded by 19.4 toyGenes on average, most of them are coded by only a few toyGenes. For instance, $1,364$ toyProteins —roughly half of them!— are coded by less than 10 toyGenes each. On the other hand, only 4 toyProteins are coded by more than 200 toyGenes each, the maximum being 235 toyGenes coding for the same toyProtein. The distribution is close to an exponential decay (Supplementary Figure 3**a**). The same happens with the perimeters, although with less skewness: each perimeter is mapped by 7.15 toyProteins on average, but the most abundant perimeters correspond to 26 toyProteins, and 100 are mapped by 1 or 2 toyProteins each (Supplementary Figure 3**b**). As we will see later, this already induces a certain degree of neutrality in ToyLIFE phenotypes.

Folding energies range from -18.0 to -0.6, with an average in -9.63. The distribution is unimodal, although very rugged (Supplementary Figure 3**c**). Note that folding energies are discrete, and that separations between them are not equal. For instance, there are 6 toyProteins that have a folding energy of -18.0, but the next energy level is -16.3, realised by 17 toyProteins, and yet the next level is -16.0, realised by 14 toyProteins. The mode of the distribution is -10.6, realised by 202 toyProteins.

We can also study the structure of the toyProtein network (Supplementary Figure 3**e**, **f**). The nodes of this network will be the $2,710$ toyProteins. toyProtein 1 and toyProtein 2 will be neighbors if there is a pair of toyGenes that express each toyProtein and whose sequence is equal but for one toyN. The weight of the edge between toyProtein1 and 2 will be the sum of such pairs of toyGenes. It is surprising that there are no self-loops in this network —there are no mutations connecting one toyProtein to itself. In other words, although there is a strong degeneracy in the mapping from toyGenes to toyProteins, there are no connected neutral networks. If we consider just the perimeters, however, the neutrality is somewhat recovered: out of the 379 perimeters, 224 of them have neutral neighbors. So there are many mutations that alter the folding energy of a toyProtein without changing the perimeter. In this sense,

[1] In [1], where we first presented ToyLIFE, we did not use this rule: whenever a sequence folded into two folds with the same folding energy and same number of Hs in the perimeter, we would discard them. This version of ToyLIFE, therefore, is slightly different. However, the results are qualitatively similar.



Supplementary Figure 3: Distributions of toyProteins in ToyLIFE. (a) Distribution of toyProtein abundances —that is, the number of toyGenes that code for them. Most toyProteins are coded by few toyGenes, but some of them are very abundant: the most abundant toyProtein is coded by 235 toyGenes. **(b)** Distribution of the perimeters associated with each toyProtein. Again, not all perimeters are equally abundant, and some of them correspond to as many as 25 toyProteins, while 100 correspond to 1 or 2 toyProteins. **(c)** Distribution of folding energies. The range of folding energies goes from -18.0 to -0.6, with a unimodal, rugged distribution. The mode is -10.6, a folding energy achieved by 202 toyProteins. **(d)** Degree distribution in the toyProtein network. Two toyProteins are connected if there are two toyGenes coding for them that have the same sequence, except for one toyN. The average degree is 32.2. **(e)** Degree distribution in the perimeter network. Two perimeters are neighbors if the toyProteins associated to them are neighbors. The average degree is 53.3.

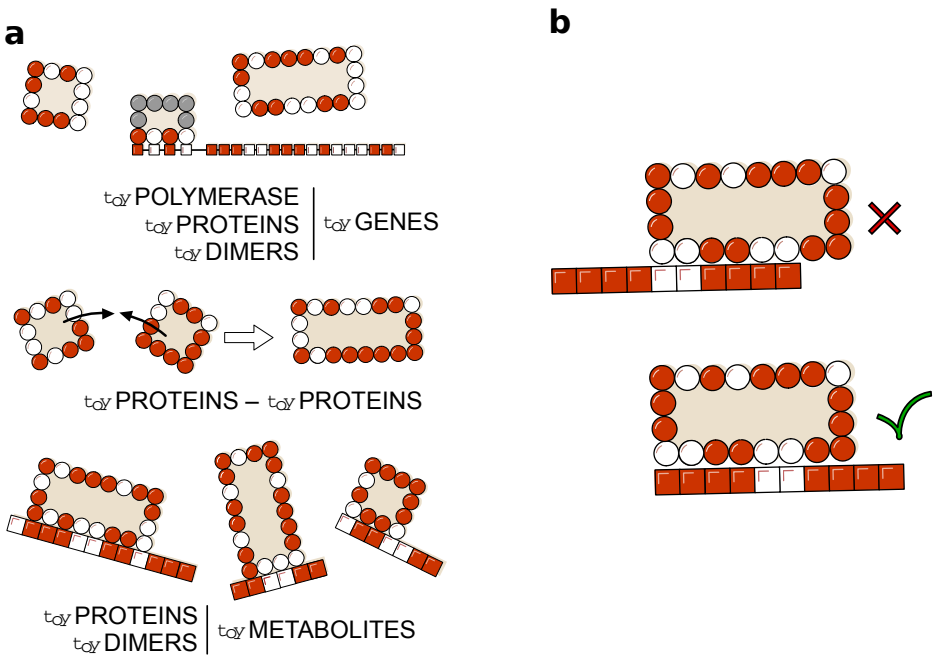
ToyLIFE is capturing a complex detail of molecular biology: mutations appear to be neutral from one point of view —in this case, perimeter— but are rarely entirely neutral. In other words, the value of a mutation is context and environment-dependent. There are always some small changes in the molecule —in this case, folding energy— that may affect their function later down the line. Real world examples of this *cryptic* effects of mutations on molecules are everywhere [4–7]. Connections between toyProteins are scarce too: the average degree in the toyProtein network is 32.2 (with a standard deviation of 25.7), a very small number — on average, each toyProtein is connected to hardly 1% of the rest of toyProteins! (Supplementary Figure 3**e**). The maximum degree is 190. This means that mutating from one toyProtein to another is not easy in general. In terms of perimeters this is more relaxed, as the average degree in the perimeter network is 53.3 (standard deviation is 38.1), with a maximum degree of 173. On average, every perimeter is connected to 14% of the rest of perimeters: it is a small number, but it is still higher than in the toyProtein case (Supplementary Figure 3**f**).

In the ToyLIFE universe, only the folding energy and perimeter of a toyProtein matter to characterise its interactions, so folded chains sharing these two features are indistinguishable. This is a difference with respect to the original HP model, where different inner cores defined different proteins and the composition of the perimeter was not considered as a phenotypic feature. However, subsequent versions of HP had already included additional traits [8].

The toyPolymerase (Supplementary Figure 1) is a special toyA polymer, similar to a toyProtein in many aspects, but that is not coded for by any toyGene. It has only one side, with sequence PHPH, and its folding energy is taken to be -11.0. We will discuss its function and place later on.

1.2 Extending the HP model: interactions

toyProteins interact through any of their sides with other toyProteins, with promoters of toyGenes, and with toyMetabolites (see Supplementary Figure 4**a**). When toyProteins bind to each other, they form a toyDimer, which is the only protein aggregate considered in ToyLIFE. The two toyProteins disappear, leaving only the toyDimer. Once formed, toyDimers can also bind to promoters or toyMetabolites through any of their sides —binding to other toyProteins or toyDimers, however, is not permitted. In all cases, the interaction energy (E_{int}) is the sum of pairwise interactions for



Supplementary Figure 4: Interactions in toyLIFE. (a) Possible interactions between pairs of toyLIFE elements. toyGenes interact through their promoter region with toyProteins (including the toyPolymerase and toyDimers); toyProteins can bind to form toyDimers, and interact with the toyPolymerase when bound to a promoter; both toyProteins and toyDimers can bind a toyMetabolite at arbitrary regions along its sequence. **(b)** When a toyDimer or toyProtein binds to a toyMetabolite with the same energy in many places, we choose the most centered binding position. If two or more binding positions have the same energy and are equally centered, then no binding occurs.

all HH, HP and PP pairs formed in the contact —these interactions follow the rules of the HP model as well. Bonds can be created only if the interaction energy between the two molecules E_{int} is lower than a threshold energy $E_{\text{thr}} = -2.6$. Note that a minimum binding energy threshold is necessary to avoid the systematic interaction of any two molecules. Low values of the threshold would lead to many possible interactions, which would increase computation times. High values would lead to very few interactions, and we would obtain a very dull model. Our choice of $E_{\text{thr}} = -2.6$ achieves a balance: the number of interactions is large enough to generate complex behaviours, as we will see later on, while at the same time keeping the universe of interactions small enough to handle computationally. If below threshold, the total energy of the resulting complex is the sum of E_{int} plus the folding energy of all toyProteins involved. The lower the total energy, the more stable the complex. When several toyProteins or toyDimers can bind to the same molecule, only the most stable complex is formed. Consistently with the assumptions for protein folding, when this rule does not determine univocally the result, no binding is produced.

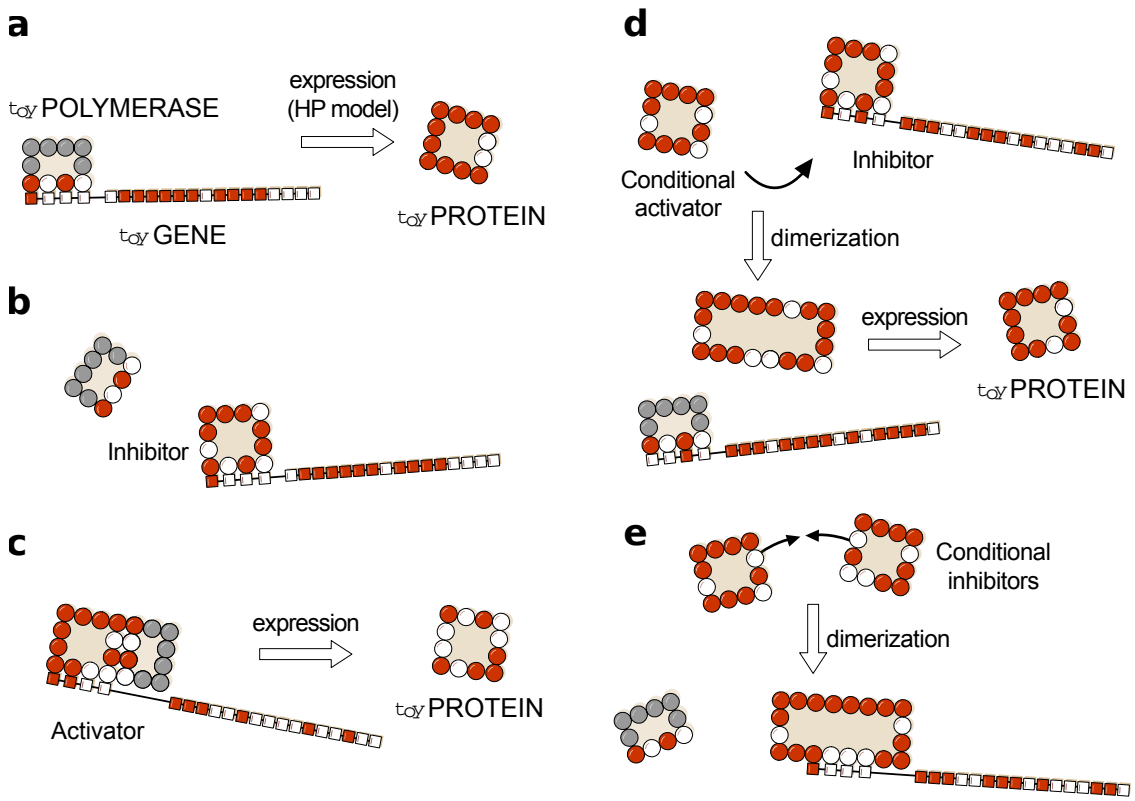
As the length of toyMetabolites is usually longer than 4 toyS (the length of interacting toyProtein sites), several binding positions between a toyMetabolite and a toyProtein might share the same energy. In those cases we select the sites that yield the most centered interaction (Supplementary Figure 4**b**). If ambiguity persists, no bond is formed. Also, no more than one toyProtein / toyDimer is allowed to bind to the same toyMetabolite, even if its length would permit it. toyProteins / toyDimers bound to toyMetabolites cannot bind to promoters.

Interaction rules in toyLIFE have been devised to remove any ambiguity. When more than one rule could be chosen, we opted for computational simplicity, having made sure that the general properties of the model remained unchanged. A detailed list of the specific disambiguation rules implemented in the model follows:

1. **Folding rule:** if a sequence of toyAminoacids can fold into two (or more) different configurations with the same energy and two different perimeters with the same number of H, it is considered degenerate and does not fold.

2. **One-side rule:** any interaction in which a toyProtein can bind any ligand with two (or more) different sides and the same energy is discarded.

3. **Annihilation rule:** if two (or more) toyProteins can bind a ligand with the same energy, the binding does not

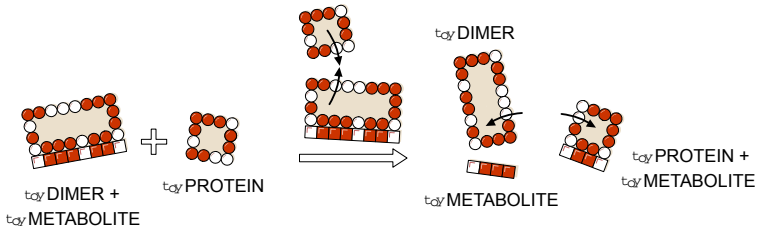


Supplementary Figure 5: Regulatory functions in toyLIFE. (a) A toyGene is expressed (translated) when the toyPolymerase binds to its promoter region. The sequence of Ps and Hs of the toyProtein will be exactly the same as that of the toyGene coding region. **(b)** If a toyProtein binds to the promoter region of a toyGene with a lower energy than the toyPolymerase does, it will displace the latter, and the toyGene will not be expressed. This toyProtein acts as an *inhibitor*. **(c)** The toyPolymerase does not bind to every promoter region. Thus, not all toyGenes are expressed constitutively. However, some toyProteins will be able to bind to these promoter regions. If, once bound to the promoter, they bind to the toyPolymerase with their rightmost side, the toyGene will be expressed, and these toyProteins act as *activators*. **(d)** More complex interactions —involving more elements— appear. For example, a toyProtein that forms a toyDimer with an inhibitor —preventing it from binding to the promoter— will effectively activate the expression of the toyGene. However, it does neither interact with the promoter region nor with the toyPolymerase, and its function is carried out only when the inhibitor is present. We call this kind of toyProteins *conditional activators*. **(e)** Two toyProteins can bind together to form a toyDimer that inhibits the expression of a certain toyGene. As they need each other to perform this function, we call them *conditional inhibitors*. As the number of genes increases, this kind of complex relationships can become very intricate.

occur. However, if a third toyProtein can bind the ligand with greater (less stable) energy than the other two, and does so uniquely, it will bind it.

4. **Identity rule:** an exception to the Annihilation rule occurs if the competing toyProteins are the same. In this case, one of them binds the ligand and the other(s) remains free.

5. **Stoichiometric rule:** an extension of the Identity rule. If two (or more) copies of the same toyProtein / toyDimer / toyMetabolite are competing for two (or more) different ligands, there will be binding if the number of copies of the toyProtein / toyDimer / toyMetabolite equals the number of ligands. For example, say that P1 binds to P2, P3 and P4 with the same energy. Then, (a) if P1, P2 and P3 are present, no complex will form; (b) if there are two copies of P1, dimers P1-P2 and P1-P3 will both form; but (c) if P4 is added, no complex will form. Conversely, if all ligands are copies as well, the Stoichiometry rule does not apply. For example, three copies of P1 and two copies of P2 will form two copies of dimer P1-P2, and one copy of P1 will remain free.



Supplementary Figure 6: Metabolism in toyLIFE. A toyDimer is bound to a toyMetabolite when a new toyProtein comes in. If the new toyProtein binds to one of the two units of the toyDimer, forming a new toyDimer energetically more stable than the old one, the two toyProteins will unbind and break the toyMetabolite up into two pieces. We say that the toyMetabolite has been catabolised.

1.3 Regulation

Expression of toyGenes occurs through the interaction with the toyPolymerase, which is a special kind of toyProtein (see Supplementary Figure 1). The toyPolymerase only has one interacting side (with sequence PHPH) and its folding energy is fixed to value -11.0: it is more stable than more than half the toyProteins. It is always present in the system. The toyPolymerase binds to promoters or to the right side of a toyProtein / toyDimer already bound to a promoter. When the toyPolymerase binds to a promoter, translation is directly activated and the corresponding toyGene is expressed (Supplementary Figure 5**a**). However, a more stable (lower energy) binding of a toyProtein or toyDimer to a promoter precludes the binding of the toyPolymerase. This inhibits the expression of the toyGene, except if the toyPolymerase binds to the right side of the toyProtein / toyDimer, in which case the toyGene can be expressed.

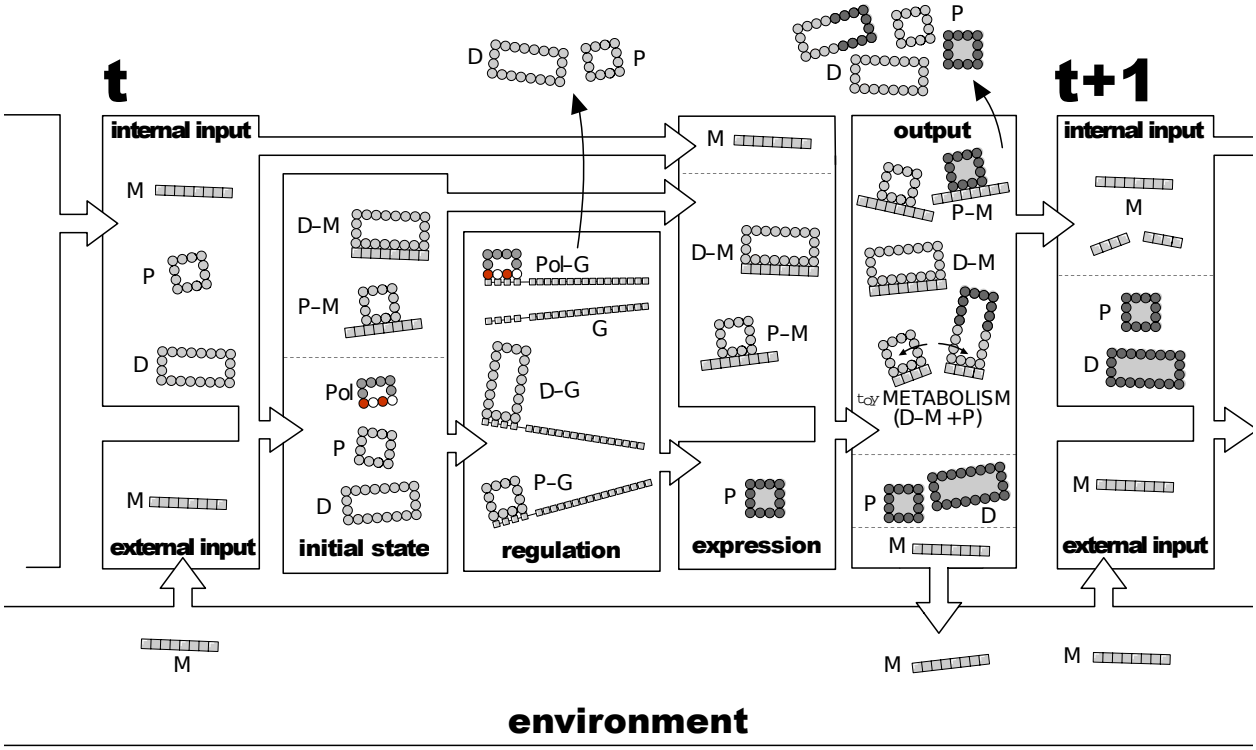
The minimal interaction rules that define toyLIFE dynamics endow toyProteins with a set of possible activities not included *a priori* in the rules of the model (see Supplementary Figure 5). For example, since the 4-toyN interacting site of the toyPolymerase cannot bind to all promoter regions —because some of these interactions have $E_{\text{int}} > E_{\text{thr}}$—, translation mediated by a toyProtein or toyDimer binding might allow the expression of genes that would otherwise never be translated. These toyProteins thus act as activators (Supplementary Figure 5**c**). This process finds a counterpart in toyProteins that bind to promoter regions more stably than the toyPolymerase does, and therefore prevent gene expression —this happens if $E_{\text{int(PROT)}} + E_{\text{PROT}} < E_{\text{int(POLY)}} + E_{\text{POLY}}$. They are acting as inhibitors (Supplementary Figure 5**b**). There are two additional functions that could not be foreseen and involve a larger number of molecules. A toyProtein that forms a toyDimer with an inhibitor —preventing its binding to the promoter— effectively behaves as an activator for the expression of the toyGene. However, it interacts neither with the promoter region nor with the toyPolymerase, and its activating function only shows up when the inhibitor is present. This toyProtein thus acts as a conditional activator (Supplementary Figure 5**d**). On the other hand, two toyProteins can bind together to form a toyDimer that inhibits the expression of a particular toyGene. As the presence of both toyProteins is needed to perform this function, they behave as conditional inhibitors (Supplementary Figure 5**e**). This flexible, context-dependent behavior of toyProteins is reminiscent of phenomena observed in real cells [9], and permits the construction of complex toyGene Regulatory Networks (toyGRNs).

1.4 Metabolism

When a toyDimer is bound to a toyMetabolite, another toyProtein can interact with this complex and break it. This reaction will take place if the toyProtein can bind to one of the subunits of the toyDimer and the resulting complex has less total energy than the toyDimer. As with the rest of interactions, the catabolic reaction will only take place if this binding is unambiguous. As a result of this reaction, the toyDimer will be broken in two: one of the pieces will be bound to the toyProtein (forming a new toyDimer), and the other one will remain free. The toyMetabolite will break accordingly: the part of it that was bound to the first subunit will stay with it, and the other part will stay with the second subunit. Note that the toyMetabolite need not be broken symmetrically: this will depend on how the toyDimer binds to it (Supplementary Figure 6).

1.5 Dynamics in toyLIFE

The dynamics of the model proceeds in discrete time steps and variable molecular concentrations are not taken into account. A step-by-step description of toyLIFE dynamics is summarised in Supplementary Figure 7. There is an initial



Supplementary Figure 7: Dynamics of toyLIFE. Input molecules at time step t are toyProteins (Ps) (including toyDimers (Ds)) and toyMetabolites, either produced as output at time step $t-1$ or environmentally supplied (all toyMetabolites denoted Ms). Ps and Ds interact with Ms to produce complexes P-M and D-M. Next, the remaining Ps and Ds and the toyPolymerase (Pol) interact with toyGenes (G) at the regulation phase. The most stable complexes with promoters are formed (Pol-G, P-G and D-G), activating or inhibiting toyGenes. P-Ms and D-Ms do not participate in regulation. Ps and Ds not in complexes are eliminated and new Ps (dark grey) are formed. These Ps interact with all molecules present and form Ds, new P-M and D-M complexes, and catabolise old D-M complexes. At the end of this phase, all Ms not bound to Ps or Ds are returned to the environment, and all Ps and Ds in P-M and D-M complexes unbind and are degraded. The remaining molecules (Ms just released from complexes, as well as all free Ps and Ds) go to the input set of time step $t+1$.

set of molecules which results from the previous time step: toyProteins (including toyDimers and the toyPolymerase) and toyMetabolites, either endogenous or provided by the environment. These molecules first interact between them to form possible complexes (see Section 1.2) and are then presented to a collection of toyGenes that is kept constant along subsequent iterations. Regulation takes place, mediated by a competition for binding the promoters of toyGenes, possibly causing their activation and leading to the formation of new toyProteins. Binding to promoters is decided in sequence. Starting with any of them (the order is irrelevant), it is checked whether any of the toyProteins / toyDimers (including the toyPolymerase) available bind to the promoter —remember that complexes bound to toyMetabolites are not available for regulation—, and then whether the toyPolymerase can subsequently bind to the complex and express the accompanying coding region. If it does, the toyGene is marked as active and the toyProtein / toyDimer is released. Then a second promoter is chosen and the process repeated, until all promoters have been evaluated. toyGenes are only expressed after all of them have been marked as either active or inactive. Each expressed toyGene produces one single toyProtein molecule. There can be more units of the same toyProtein, but only if multiple copies of the same toyGene are present.

toyProteins / toyDimers not bound to any toyMetabolite are eliminated in this phase. Thus, only the newly expressed toyProteins and the complexes involving toyMetabolites in the input set remain. All these molecules interact yet again, and here is where catabolism can occur. Catabolism happens when, once a toyMetabolite-toyDimer complex is formed, an additional toyProtein binds to one of the units of the toyDimer with an energy that is lower than that of the initial toyDimer. In this case, the latter disassembles in favor of the new toyDimer, and in the process the toyMetabolite is broken, as already mentioned in Section 1.4 and Supplementary Figure 6. The two pieces of

the broken toyMetabolites will contribute to the input set at the next time step, as will free toyProteins / toyDimers. However, toyProteins / toyDimers bound to toyMetabolites disappear in this phase —they are degraded—, and only the toyMetabolites are kept as input to the next time step. Unbound toyMetabolites are returned to the environment. This way, the interaction with the environment happens twice in each time step: at the beginning and at the end of the cycle.

2 A note on toyMetabolites

There are 2^m binary strings —toyMetabolites— of length m. From lengths 4 to 8, therefore, there are

$$\sum_{m=4}^{8} 2^m = 496$$

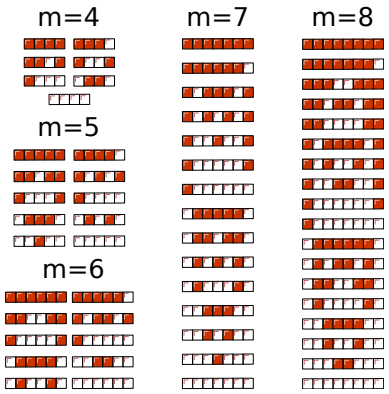
toyMetabolites. However, due to the interaction rules of tOYLIFE, a particular string and its reverse —i.e. HPPHPPPP and PPPPHPPH— will be treated the same way by tOYLIFE organisms. Therefore, for all practical purposes, we will consider each string and its reverse as the same toyMetabolite, thus staying with 274 of them. Additionally, there are 60 toyMetabolites that cannot be catabolised in tOYLIFE (Supplementary Figure 8). For all lengths, toyMetabolites formed by all Ps and one H at one extreme, or all Hs and one P at one extreme, are unbreakable. This is because there is no unambiguous way in which a toyDimer can bind to these toyMetabolites. There are two of these toyMetabolites for each length, making a total of 10. Additionally, the toyMetabolite PPHP cannot be broken due to the same reason. Symmetrical toyMetabolites, in general, cannot be catabolised either. Because of the interaction rules described in Section 1, only symmetrical toyDimers can bind to these toyMetabolites. But symmetrical toyDimers cannot be broken: any toyProtein that can bind to one subunit will be able to bind the other one. Because of the disambiguation rules, no binding is produced, and catabolism does not occur. There are 52 symmetric toyMetabolites —because they repeat half the sequence, there are

$$\sum_{m=4}^{8} 2^{\left[\frac{m+1}{2}\right]} = 52$$

of them, $[x]$ being the integer part of x —odd-length symmetrical toyMetabolites repeat $m+1$ toySugars, hence the $[(m+1)/2]$ exponent. However, three symmetrical toyMetabolites of length 7 —namely, PPPHPPP, PPHPHPP and PPHHHPP— can actually be broken. So there are 49 unbreakable symmetrical toyMetabolites. Added to the previous 11 unbreakable toyMetabolites, we get the total of 60. As a result, the total number of toyMetabolites up to length 8 is 214.

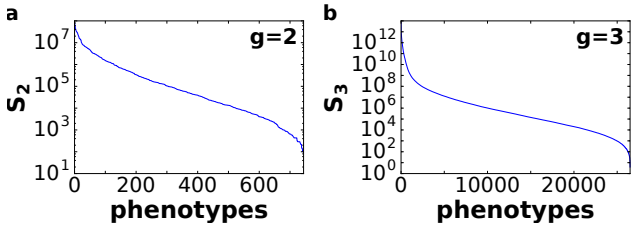
It is somewhat interesting that, as an emergent property of the model, some toyMetabolites are not able to be catabolised. Moreover, it is not that these toyMetabolites are irrelevant to the model: if they are present, they will interact with symmetric toyDimers, affecting the regulatory output of cells. So these toyMetabolites could function as signalling molecules.

What happens with longer toyMetabolites? Because of the way interactions have been defined in tOYLIFE, longer toyMetabolites can be considered as unions of shorter ones. For instance, a toyMetabolite of length 9 is (in terms of interactive potential) equal to two toyMetabolites of length 8. If a genotype is able to catabolise one of these, it will be able to catabolise the longer one, so the metabolic phenotype for toyMetabolites of arbitrary length is uniquely determined by considering lengths up to 8 toySugars.



Supplementary Figure 8: Unbreakable toyMetabolites. There are 60 unbreakable toyMetabolites: 49 of them are symmetrical, other 10 are chains of all Hs or all Ps in a row, and the last one is PPHP. Because of the interaction rules in tOYLIFE, only symmetrical toyDimers would be able to bind these toyMetabolites, and therefore they cannot be broken.

3 Rank plots for phenotypes in $g = 2$ and $g = 3$



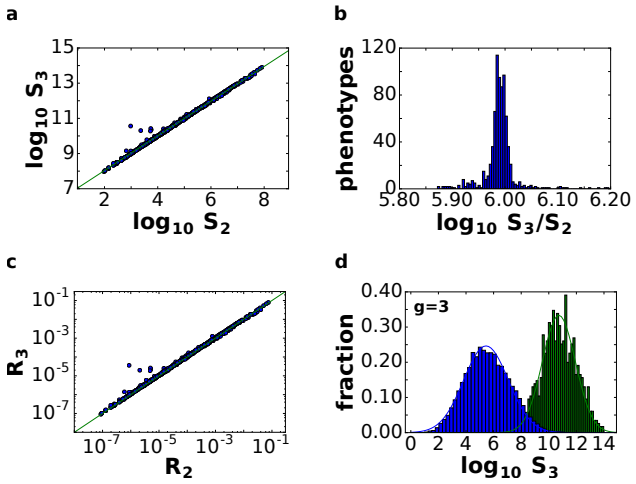
Supplementary Figure 9: Phenotype frequencies vary enormously in toyLIFE. Rank plots for all phenotypes in $g = 2$ **(a)** and $g = 3$ **(b)**. Both plots show a long tail of small phenotypes. In particular, for $g = 3$, only 300 phenotypes in $\mathcal{P}_3$ represent almost 99% of all genotypes. The remaining $26,000$ phenotypes are extremely rare by comparison.

4 Comparison between the $g = 2$ and $g = 3$ case

In Supplementary Figure 10**a**, we represent the abundance of a phenotype for $g = 2$ (S_2) versus its corresponding abundance for $g = 3$ (S_3), for each phenotype in $\mathcal{P}_2$. The Figure also shows a power-law fit, $\log_{10} S_3 = 6.064 + 0.986 \log_{10} S_2$, corresponding to $S_3 = 10^{6.064} S_2^{0.986} \approx 10^6 S_2$, a linear fit. This means that the abundance ordering between these phenotypes does not change when exploring genotypes with one more gene. The goodness of the fit is further shown in Supplementary Figure 10**b**, which represents the histogram of values of $\log_{10}(S_3/S_2)$. The distribution is concentrated around its mean, 5.996, very close to the value 6.064 obtained in Supplementary Figure 10**a**. This second result confirms that the abundance of $\mathcal{P}_2$ phenotypes in $g = 3$ space is equal to their corresponding abundance in $g = 2$ space times 10^6. Where does this factor come from? Recall that there are $2^{20} \sim 10^6$ toyGenes in toyLIFE. A factor of almost 10^6 between S_3 and S_2 means that we can add almost any toyGene to a given two-gene genotype, and the resulting phenotype will be the same: it will not interfere with the original function. This is a remarkable fact.

Moreover, if we look at the distribution of relative abundances of $\mathcal{P}_2$ phenotypes —computed as phenotype abundance divided by the total number of viable genotypes— for $g = 2$ and $g = 3$ (Supplementary Figure 10**c**), we obtain a linear relationship again: $R_3 = R_2$. Which means that the relative abundance of the phenotypes for $g = 2$ is very similar to the relative abundance they represent for $g = 3$. But the sum of the relative abundances for $g = 2$ is equal to 1 —there are only 775 phenotypes in $\mathcal{P}_2$. Accordingly, the sum of relative abundances for $g = 3$ is close to 1 —actually, it is 0.9964.

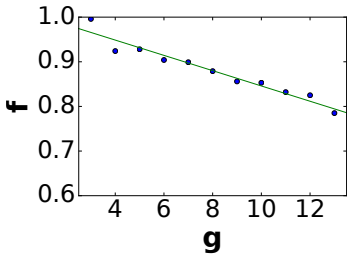
Finally, let us look again at the histogram of phenotype abundance distributions in $g = 3$ that we obtained in Figure 1**c** (main text). We can re-compute the histogram taking the 775 phenotypes from $\mathcal{P}_2$ as a separate set from the remaining $25,717$ phenotypes in $\mathcal{P}_3 - \mathcal{P}_2$. If we compute the respective histograms for both sets, we obtain Supplementary Figure 10**d**. In green we have represented the 775 phenotypes in $\mathcal{P}_2$. It is not surprising that their distribution follows a log-normal law again: it follows immediately from Figure 1**a** (main text) and from the linear relationship shown in Supplementary Figure 10**a**. What is relevant, however, is that the *bump* we observed in Figure 1**c** (main text) is gone in the histogram of the remaining $25,717$ phenotypes (in blue).



Supplementary Figure 10: Two-gene phenotypes dominate phenotype space in the three-gene case. (a) The 775 phenotypes belonging to $\mathcal{P}_2$ also appear in $\mathcal{P}_3$. This figure represents the corresponding abundance of each phenotype in both genotype spaces: S_2 and S_3 are, respectively, phenotype abundance for $g = 2$ and $g = 3$. Green line represents the linear fit $\log_{10} S_3 = 6.064 + 0.986 \log_{10} S_2$, which is close to the linear fit $S_3 \sim 10^6 S_2$. **(b)** Histogram of $\log_{10}(S_3/S_2)$ for each of the 775 phenotypes in $\mathcal{P}_2$. The mean of the distribution is 5.996. **(c)** Relative abundance of the 775 phenotypes in $\mathcal{P}_2$ (R_2) versus their relative abundance for $g = 3$ (R_3) —computed as phenotype abundance divided by number of viable genotypes. Green line is $R_3 = R_2$. The close fit means that the phenotypes from $\mathcal{P}_2$ dominate phenotype space for $g = 3$. **(d)** Abundance distribution of phenotypes in $\mathcal{P}_3$, taking the 775 phenotypes in $\mathcal{P}_2$ and rescaling them — we have obtained the two histograms as if they came from independent distributions for clarity. The green histogram represents the phenotypes in $\mathcal{P}_2$, and the blue histogram the remaining $25,717$ phenotypes in $\mathcal{P}_3$. New log-normal fits are drawn: $\mu_3 = 5.449$, $\sigma_3 = 1.619$ (blue line), $\mu_2 = 10.730$, $\sigma_2 = 1.196$ (green line). Note that the log-normal fit for three-gene phenotypes is much better once we take into account the 775 phenotypes in $\mathcal{P}_2$. All fits in this and subsequent Supplementary Figures have been done using the least squares method.

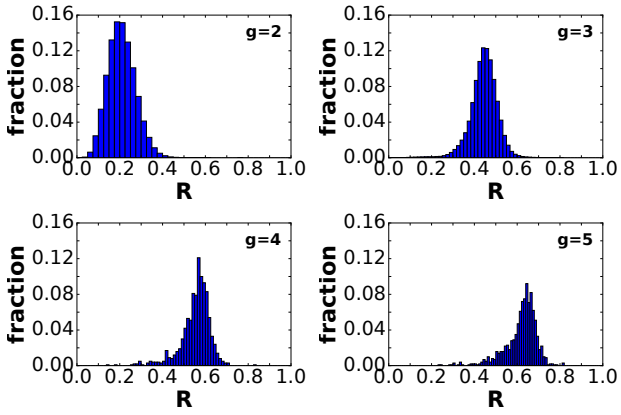
5 Relevance of $\mathcal{P}_2$ phenotypes

A relevant question now is how important the 775 phenotypes in $\mathcal{P}_2$ are for larger genotypes. Exhaustive sampling of genotype spaces larger than $g = 3$ is out of our possibilities, but we can perform random samples of genotypes for different values of g and observe the fraction f of observed phenotypes that belong to $\mathcal{P}_2$. This is represented in Supplementary Figure 11. Observe that, although this fraction decays linearly with gene size as $f = 1.02 - 0.02g$, the slope of the decay is very small, and therefore the fraction is always high —higher than 80% for $g \leq 13$. In other words, phenotypes in $\mathcal{P}_2$ continue to dominate phenotype space in tOYLIFE for a moderate number of genotype sizes.



Supplementary Figure 11: The dominance of two-gene phenotypes decays linearly with genotype size. For each g, we sample $10,000$ viable genotypes and compute their phenotypes, counting how many phenotypes belong to $\mathcal{P}_2$. We then represent the fraction f versus g. The data can be fitted to a linear function: $f = 1.02 - 0.02g$ (green line). The fraction of phenotypes belonging to $\mathcal{P}_2$ decays with g, albeit very slowly.

6 Robustness histograms in toyLIFE



Supplementary Figure 12: Genotypes in toyLIFE typically have a large number of neutral neighbors. Distribution of robustness for genotypes for different values of g (gene number) for $g = 2$ to $g = 5$. Robustness is defined as the normalized degree of a node in the networks: $R = k/k_{\max}$, where k is the degree of a node in the neutral network, and $k_{\max} = 20g$ is the maximum degree in the network. Normalisation allows us to compare values for different genotype sizes. For $g = 2$ and $g = 3$, we sampled 10^7 genotypes, whereas for $g = 4$ and $g = 5$ we sampled $1,000$ genotypes. All distributions are unimodal, and more or less concentrated around the mean.

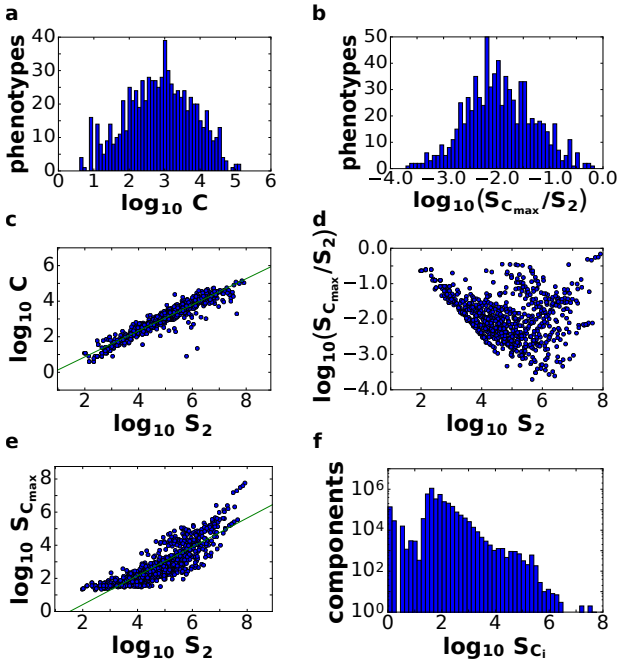
7 Connected components for $g = 2$

For $g = 2$ we can perform network analyses on all 775 phenotypes exhaustively, and compute their connected components (Supplementary Figure 13). We observe that most phenotypes are distributed in highly fragmented neutral networks: the genotypes corresponding to a given phenotype cluster in many disjoint connected components (Supplementary Figure 13**a**): the number of connected components C is never smaller than 4 and is usually much larger. Moreover, these connected components tend to be small: if we consider Cmax, the maximal component associated to each neutral network, its average relative abundance $S_{C\text{max}}/S_2$ is 0.033 (Supplementary Figure 13**b**). Only 63 phenotypes have connected components that are larger than 10% the phenotype abundance —among these are the largest connected components for $g = 2$, including one giant network that contains $56,889,472$ nodes!

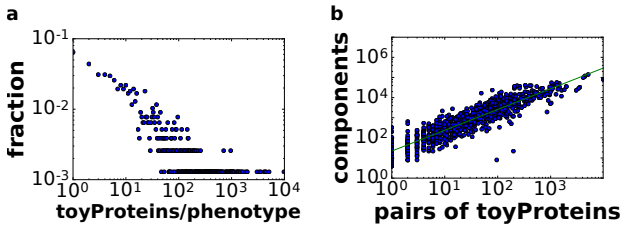
Large phenotypes tend to have a larger number of connected components, and we can find a relatively good power-law fit between the abundance of the phenotype S_2 and the number of components C: $C = 0.25S_2^{0.7}$ (Supplementary Figure 13**c**). The relationship between S_2 and the relative size of Cmax is noisy (Supplementary Figure 13**d**): smaller phenotypes have less connected components and therefore the relative size of the maximal component is high. As the number of components increases, most of them tend to have equal, small sizes. However, the largest phenotypes with the greatest number of connected components also have the largest connected components, as we pointed out before, so there is a positive correlation between S_2 and the absolute size of its maximal component, $S_{C\text{max}}$. This last fact is represented in Supplementary Figure 13**e**.

In short, there is a huge variation in the size of connected components in $g = 2$. We can plot the distribution of sizes of all connected components C_i —irrespective of the phenotype they belong to (Supplementary Figure 13**f**). The average component size, S_{C_i}, is 301.4, but we can see from the histogram that the distribution has a long tail. Therefore, although most connected components are smaller than $1,000$ nodes — roughly 98.5%!— some of them reach up to $\sim 10^7$ nodes.

The high disconnection in connected components is due to the HP model that underlies toyProtein folding. Any



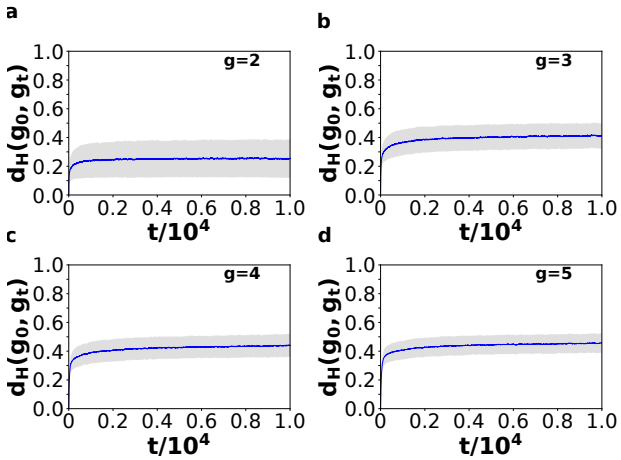
Supplementary Figure 13: Neutral networks in toyLIFE are highly fragmented for $g = 2$. (a) For all 775 phenotypes in $\mathcal{P}_2$, we computed the number of connected components (C) of the associated neutral network. This figure represents the distribution of the decimal logarithm of C per neutral network. No single phenotype has less than 4 connected components. **(b)** For each neutral network, we take the maximal component Cmax and plot the distribution of the logarithm of its relative size —that is, the logarithm of $S_{C\text{max}}$ divided by S_2. **(c)** The abundance of the phenotype and the number of components are related via a power law: $C = 0.25S_2^{0.7}$. **(d)** The relationship between the relative abundance of Cmax and the abundance of the phenotype is very noisy, but **(e)** there is a positive correlation between the absolute abundance of Cmax and the abundance of the phenotype. The green line represents the power law fit $S_{C\text{max}} = 0.05_2^{0.9}$. **(f)** Distribution of the logarithm of abundance of all connected components C_i for $g = 2$.



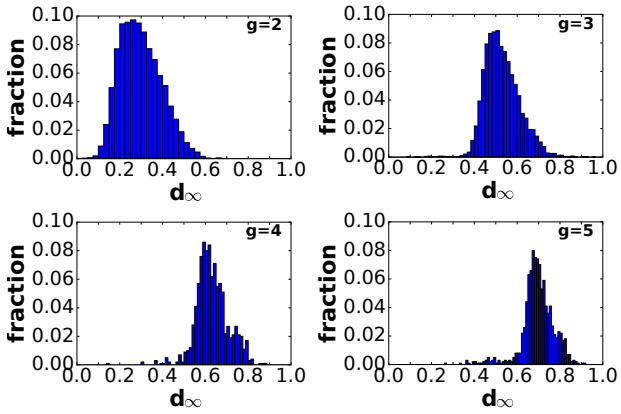
Supplementary Figure 14: Most phenotypes in $\mathcal{P}_2$ are obtained by a small number of pairs of toyProteins. (a) Distribution of the number of pairs of toyProteins that generate a given phenotype. For example, if both $\{1,1\}$, $\{1,2\}$ and $\{3,4\}$ generate a given phenotype, there are 3 pairs of toyProteins that generate it. **(b)** Due to the HP model that underlies toyProtein folding, the more pairs of toyProteins are able to generate a given phenotype, the larger the phenotype and, because of the power-law relationship obtained in Supplementary Figure 13**c**, the more connected components that will belong to the phenotype. The green line represents the power-law fit $C = 22.093P^{1.032}$.

given phenotype in $\mathcal{P}_2$ will be obtained by some set of pairs of toyProteins. Supplementary Figure 14**a** shows that this distribution is highly skewed, with a long tail: 28.64% of phenotypes in $\mathcal{P}_2$ are obtained by less than 10 pairs of toyProteins, while one phenotype is obtained by $9,808$ pairs of toyProteins. The problem, therefore, is not due to a small set of toyProteins associated to each phenotype. Rather, the cause of the disconnection between connected components is due to the lack of neutral mutations among proteins and the difficulty to reach different proteins in toyLIFE (see main text).

8 Random walks in toyLIFE



Supplementary Figure 15: Neutral networks in toyLIFE span a large fraction of genotype space (1). For each genotype size, from $g = 2$ to $g = 5$, we performed $1,000$ neutral random walks starting at randomly chosen genotypes. The length of the random walks was $10,000$ time steps. The figure represents the average Hamming distance $\langle d_H \rangle$ (blue line) between the genotype visited at time t, g_t, and the original genotype g_0, plus minus one standard deviation (grey area), empirically obtained from the data.

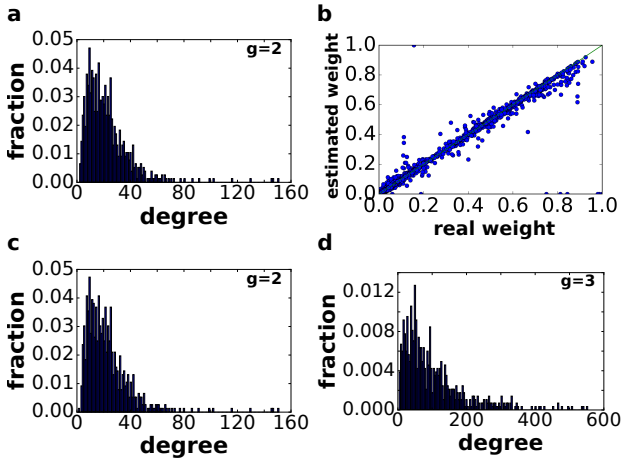


Supplementary Figure 16: Neutral networks in toyLIFE span a large fraction of genotype space (2). Distribution of d_∞ for genotypes with different values of g (gene number) for $g = 2$ to $g = 5$. We performed $10,000$ (for $g \leq 3$) or $1,000$ (for $g > 3$) neutral random walks, forcing them to increase the Hamming distance to the original genotype. We stopped when the random walk could not get farther.

9 Connections between phenotypes

For $g = 2$ we can build the phenotype network exhaustively. The network is not entirely connected: there is a giant component that includes 767 nodes out of the 775, and six additional tiny components, five of them with just one node and the remaining one with three nodes. Additionally, the results show that the average degree is low, just 22.1, with a standard deviation of 17.3 (Supplementary Figure 17**a**). The maximum degree is 151 and the minimum is 2. The largest weights are always those of the self-loops —that is, the majority of connections in the genotype network do not change phenotype, consistently with our previous discussion on robustness. In fact, because not all phenotypes are equally large, we can compute the weighted average degree of the network —giving more weight to larger phenotypes. The result is an average degree of 54.0, illustrating that larger phenotypes are more connected than the average.

For $g = 3$, we cannot build the phenotype network exhaustively. We will resort to a numerical approximation, in order to estimate the degrees of the nodes and their relative weights. Suppose we perform a random walk over all viable genotypes —jumping among them without any additional rule. If all genotypes are connected to each other —given our results for $g = 2$, this does not seem a terrible assumption— then we expect that, as the length of the random walk tends to infinity, every phenotype is visited proportionally to its abundance, and that the visits from one phenotype to another are proportional to the actual number of connections between them. The average number of visits (per time step) from phenotype i to j as time tends to infinity will be the same as the number of connections between phenotypes i and j, divided by the total number of connections leaving i. We can check if this approach is accurate by performing the random walk for $g = 2$, for which we have the actual connection data. We performed a random walk starting at a randomly chosen genotype for 10^9 time steps. The relative weights computed by this method are close to the actual weights, as shown in Supplementary Figure 17**b**. The correlation between both variables is 0.978: the outliers correspond to small phenotypes, which are hardly visited in the random walk. Supplementary Figure 17**c** shows that the estimated degree distribution is very similar to the actual one (Supplementary Figure 17**a**). Having made sure that this approach works, we repeated it for $g = 3$, again with a random walk of length 10^9 time steps. We restricted the random walk to the 775 phenotypes in $\mathcal{P}_2$: we wanted to study how the addition of one gene altered the connections between these phenotypes. When one mutation left this set of phenotypes, we considered it as lethal. The results obtained show that all phenotypes in $\mathcal{P}_2$ now belong to one giant component —there is one phenotype that does not appear in the sample, but did belong to the giant component in $g = 2$, so it must belong to it in $g = 3$. The average degree is higher, 101.1, with a standard deviation of 90.3 (Supplementary Figure 17**d**). The maximum degree is 553, and the minimum is 4. The degree distribution is much wider, and the connectivity between phenotypes has been greatly enhanced. The weighted average degree is 333.3, again showing that larger phenotypes are much more connected than smaller ones.



Supplementary Figure 17: Connections between phenotypes in toyLIFE. (a) Degree distribution of the phenotype network in $g = 2$. Two phenotypes are connected if there is at least one genotype belonging to the first that can mutate into another genotype belonging to the second phenotype. The average degree is 22.134. **(b)** Estimated relative weight between phenotypes versus actual relative weight. Estimation performed by a random walk among all viable genotypes in $g = 2$. Length of the random walk is 10^9. The correlation between both variables is 0.978. **(c)** Estimated degree distribution from the previous random walk, for $g = 2$. **(d)** Estimated degree distribution for $g = 3$, using a random walk among genotypes belonging to phenotypes in $\mathcal{P}_2$.

References

[1] Arias CF, Catalán P, Manrubia S, Cuesta JA. toyLIFE: a computational framework to study the multi-level organisation of the genotype-phenotype map. Sci Rep. 2014;4:7549.

[2] Li H, Helling R, Tang C, Wingreen N. Emergence of preferred structures in a simple model of protein folding. Science. 1996;273:666–669.

[3] Dill KA. Theory for the folding and stability of globular proteins. Biochemistry. 1985;24:1501–1509.

[4] Aharoni A, Gaidukov L, Khersonsky O, Gould SM, Roodveldt C, Tawfik DS. The'evolvability'of promiscuous protein functions. Nat Genet. 2005;37:73–76.

[5] Amitai G, Gupta RD, Tawfik DS. Latent evolutionary potentials under the neutral mutational drift of an enzyme. HFSP J. 2007;1:67–78.

[6] Khersonsky O, Tawfik DS. Enzyme promiscuity: a mechanistic and evolutionary perspective. Ann Rev Biochem. 2010;79:471–505.

[7] Hayden EJ, Ferrada E, Wagner A. Cryptic genetic variation promotes rapid evolutionary adaptation in an RNA enzyme. Nature. 2011;474:92–95.

[8] Hoque T, Chetty M, Sattar A. Extended HP model for protein structure prediction. J Comput Biol. 2009;16:85–103.

[9] Piatigorsky J. Gene Sharing and Evolution: the Diversity of Protein Functions. Harvard University Press Cambridge MA:; 2007.