toyLIFE

Pablo Catalán, Clemente F. Arias, Susanna Manrubia and José A. Cuesta

Contents

toyLIFE was originally presented in [1]. We give here its main details, with slight modifications in the definition of the model, as presented in [2].

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 (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 [3, 4].

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 SAWs on that lattice (Figure 2).

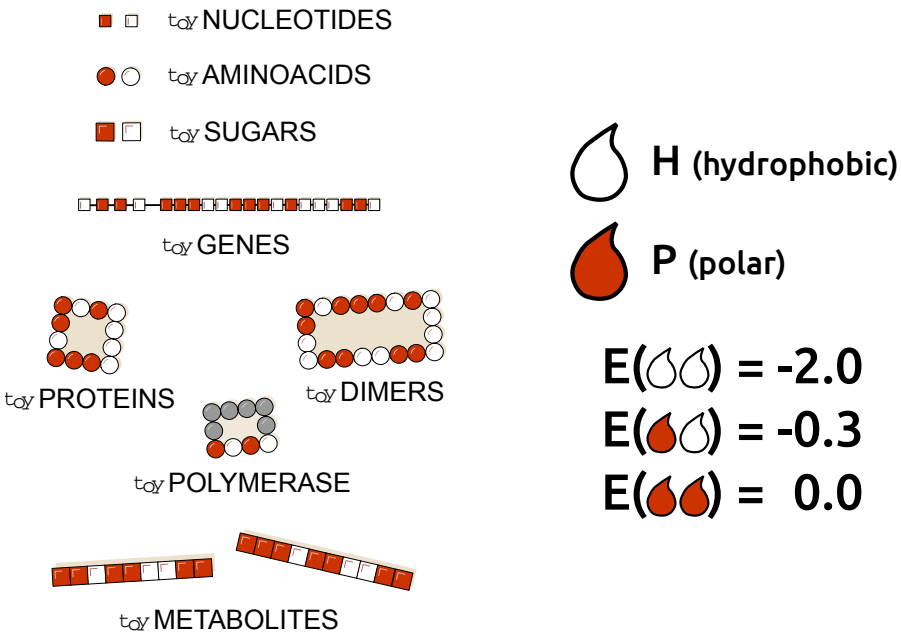

Figure 1: Building blocks and interactions defining toyLIFE. The three basic building blocks of toyLIFE are toyNucleotides, toyAminoacids, and toySugars. They can be hydrophobic (H, white) or polar (P, red), and their random polymers constitute toyGenes, toyProteins, and toyMetabolites. The toyPolymerase 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 $E_{\mathsf{HH}} = -2$, $E_{\mathsf{HP}} = -0.3$ and $E_{\mathsf{PP}} = 0$.

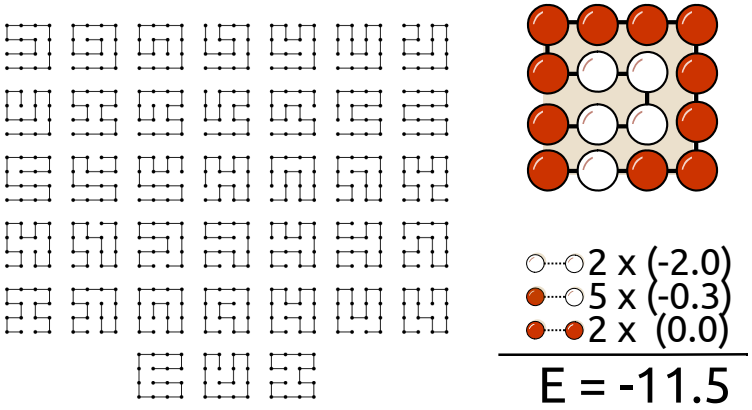

Figure 2: Protein folding in $\mathtt{toy}$LIFE. toyProteins fold on a 4×4 lattice, following a self-avoiding walk (SAW). Discarding symmetries, there are 38 SAWs (left). For each binary sequence of length 16, we fold it into every SAW and compute its folding energy, following the HP model. For instance, 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.

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 $E_{\mathsf{HH}} = -2$, $E_{\mathsf{HP}} = -0.3$ and $E_{\mathsf{PP}} = 0$, following the conditions set in [4] that $E_{\mathsf{PP}} > E_{\mathsf{HP}} > E_{\mathsf{HH}}$ (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 [2]. 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. As we will see later, this already induces a certain degree of neutrality in $\mathtt{toy}$LIFE phenotypes.

Folding energies range from -18.0 to -0.6, with an average in -9.63. The distribution is unimodal, although very rugged [2]. 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.

There are no neutral mutations at the toyProtein level in $\mathtt{toy}$LIFE. Mutations will lead to a degenerate toyPreotein that does not fold or, less often, to a different toyProtein. 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, $\mathtt{toy}$LIFE 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.

In the $\mathtt{toy}$LIFE universe, only the folding energy and perimeter of a toyProtein matter to characterise its interactions,

[1] In [1], where we first presented $\mathtt{toy}$LIFE, 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 $\mathtt{toy}$LIFE, therefore, is slightly different. However, the results are qualitatively similar.

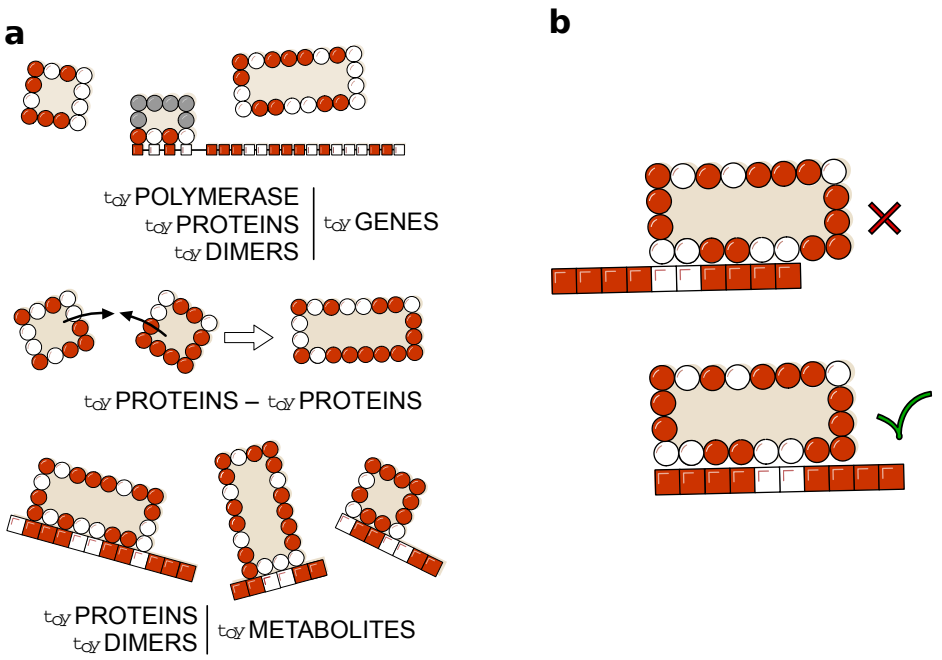

Figure 3: Interactions in toyLIFE. (a) Possible interactions between pairs of toyLIFE elements. toyGenes interact through their promoter region with toyProteins (including the toyPolymerase and toyDimers); toyProteins can bind to form toyDimers, and interact with the toyPolymerase when bound to a promoter; 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.

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.

The toyPolymerase (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.

2 Extending the HP model: interactions

toyProteins interact through any of their sides with other toyProteins, with promoters of toyGenes, and with toyMetabolites (see Figure 3**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 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 (Figure 3**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 $\mathtt{toy}$LIFE 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 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.

3 Regulation

Expression of toyGenes occurs through the interaction with the toyPolymerase, which is a special kind of toyProtein (see 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 (Figure 4**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 $\mathtt{toy}$LIFE dynamics endow toyProteins with a set of possible activities not included *a priori* in the rules of the model (see Figure 4). 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 (Figure 4**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 (Figure 4**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 (Figure 4**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 (Figure 4**e**). This flexible, context-dependent behavior of toyProteins is reminiscent of phenomena observed in real cells, and permits the construction of complex toyGene Regulatory Networks (toyGRNs).

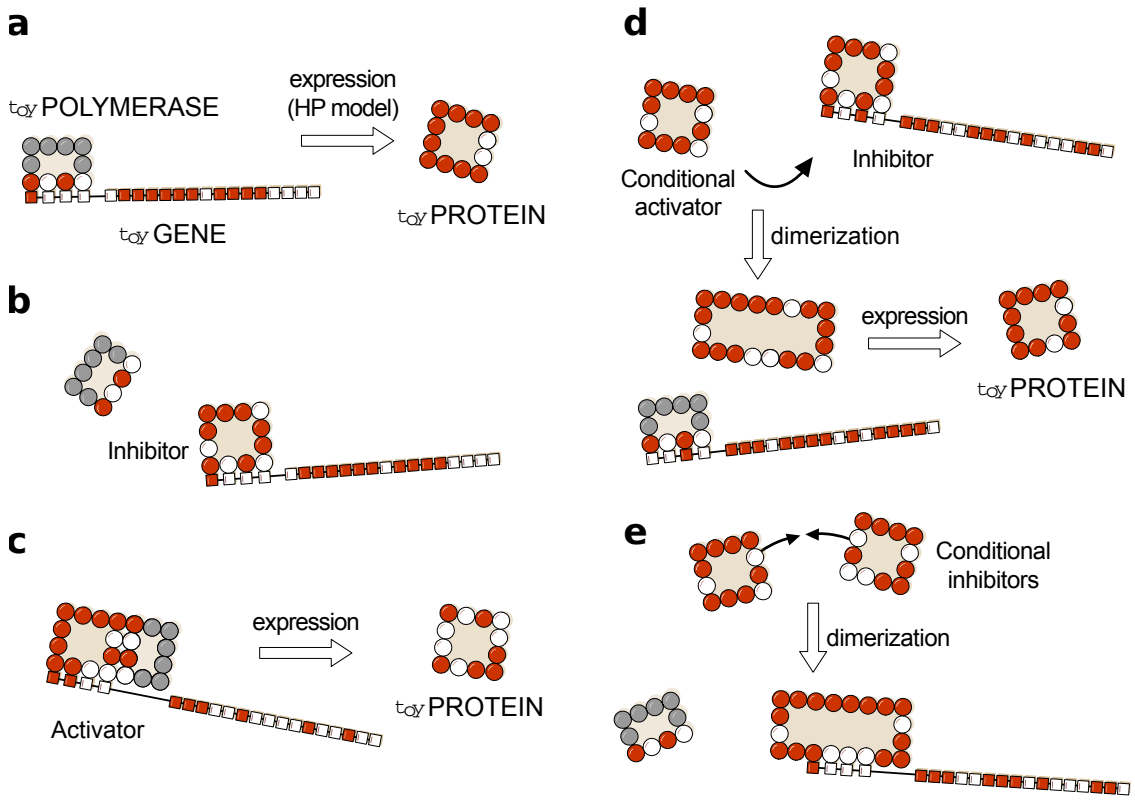

Figure 4: Regulatory functions in toyLIFE. (a) A toyGene is expressed (translated) when the toyPolymerase binds to its promoter region. The sequence of Ps and Hs of the toyProtein will be exactly the same as that of the toyGene coding region. **(b)** If a toyProtein binds to the promoter region of a 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.

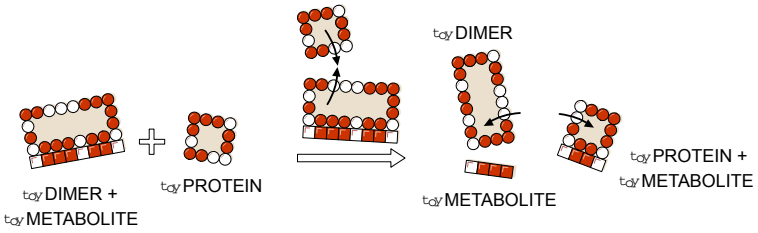

Figure 5: Metabolism in toyLIFE. A toyDimer is bound to a toyMetabolite when a new toyProtein comes in. If the new toyProtein binds to one of the two units of the toyDimer, forming a new toyDimer energetically more stable than the old one, the two toyProteins will unbind and break the toyMetabolite up into two pieces. We say that the toyMetabolite has been catabolised.

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 (Figure 5).

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 Figure 6. There is an initial 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 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 4 and Figure 5. 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.

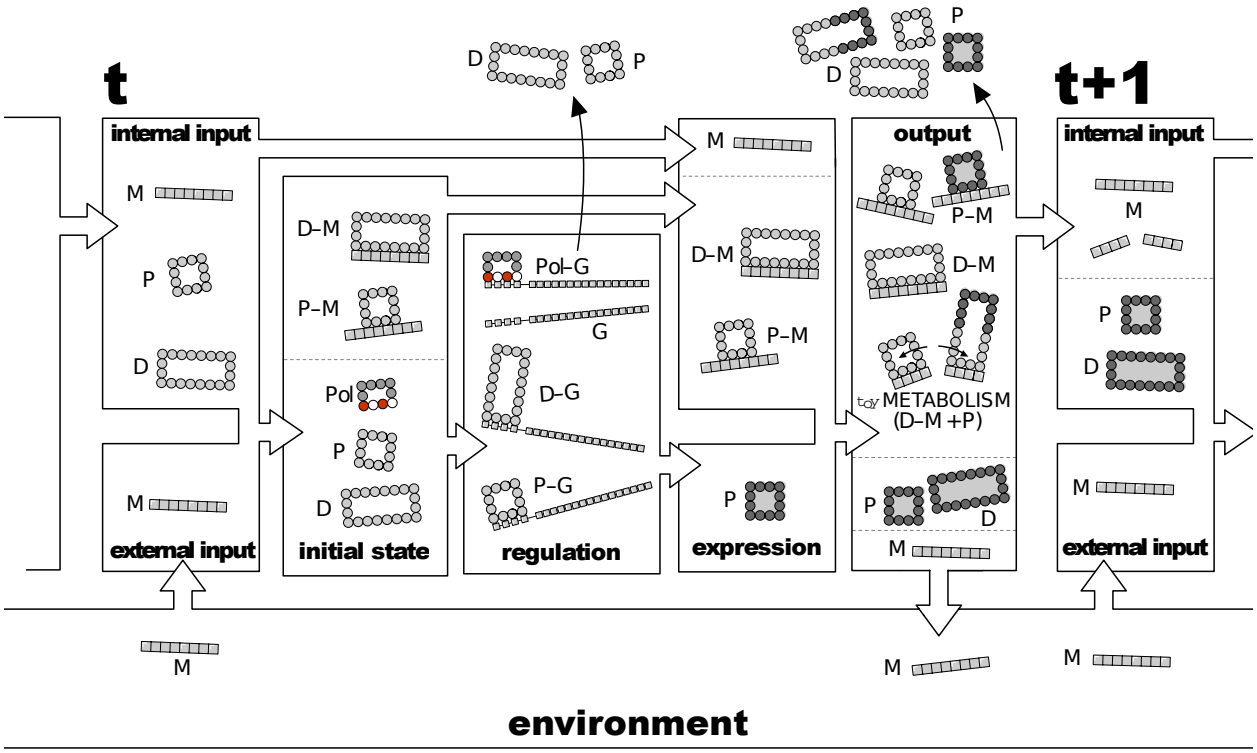

Figure 6: Dynamics of toyLIFE. Input molecules at time step t are toyProteins (Ps) (including toyDimers (Ds)) and toyMetabolites, either produced as output at time step $t-1$ or environmentally supplied (all toyMetabolites denoted Ms). Ps and Ds interact with Ms to produce complexes P-M and D-M. Next, 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 $t+1$.

6 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 $\mathtt{toy}$LIFE, a particular string and its reverse —i.e. HPPHPPPP and PPPPHPPH— will be treated the same way by $\mathtt{toy}$LIFE 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 $\mathtt{toy}$LIFE (Figure 7). 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 2, 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 $\mathtt{toy}$LIFE, 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.

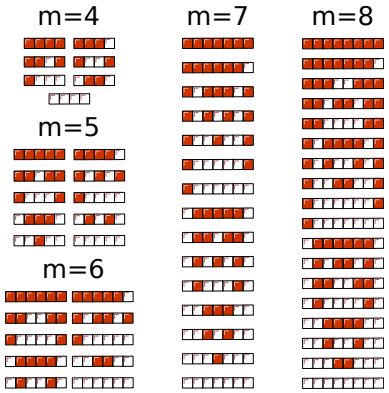

Figure 7: Unbreakable toyMetabolites. There are 60 unbreakable toyMetabolites: 49 of them are symmetrical, other 10 are chains of all Hs or all Ps in a row, and the last one is PPHP. Because of the interaction rules in $\mathtt{toy}$LIFE, only symmetrical toyDimers would be able to bind these toyMetabolites, and therefore they cannot be broken.

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] Catalán P, Wagner A, Manrubia S, Cuesta JA. Adding levels of complexity enhances robustness and evolvability in a multi-level genotype-phenotype map. J Roy Soc Interface. 2018;15:20170516.

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

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