Max-Planck-Institut für Mathematik in den Naturwissenschaften Leipzig

Network function shapes network structure: the case of the Arabidopsis flower organ specification genetic network

by

Adrien Henry, Françoise Monéger, Areejit Samal, and Olivier Martin

Preprint no.: 37 2013

Network function shapes network structure: the case of the *Arabidopsis* flower organ specification genetic network

Adrien Henry[ab], Françoise Monéger[c], Areejit Samal*[ad] and Olivier C. Martin*[ab]

[a] Laboratoire de Physique Théorique et Modèles Statistiques, CNRS UMR 8626, Université Paris-Sud, 91405 Orsay Cedex, France
[b] UMR de Génétique Végétale du Moulon, UMR 0320/UMR 8120, INRA / CNRS / Université Paris-Sud, 91190 Gif-sur-Yvette, France.
* E-mail: *olivier.martin@u-psud.fr*; Fax: +33 16933 2340
[c] Laboratoire Reproduction et Developpement des Plantes, UMR 5667, ENS / CNRS / INRA / Univ. Lyon I, 69364 Lyon cedex 07, France
[d] Max Planck Institute for Mathematics in the Sciences, Inselstr. 22, 04103 Leipzig, Germany.
* E-mail: samal@mis.mpg.de; Fax: +49 341 9959 658

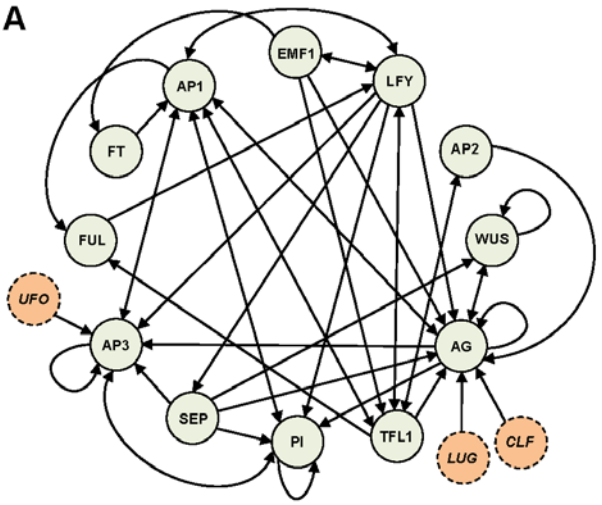
The reconstruction of many biological networks has allowed detailed studies of their structural properties. Several features exhibited by these networks have been interpreted to be the result of evolutionary dynamics. For instance the degree distributions may follow from a preferential attachment of new genes to older ones during evolution. Here we argue that even in the absence of any evolutionary dynamics, the presence of atypical features may follow from the fact that the network implements certain functions. To examine this *network function shapes network structure* scenario, we focus on the *Arabidopsis* genetic network controlling early flower organogenesis in which gene expression dynamics has been modelled using a Boolean framework. Specifically, for a system with 15 master genes, the phenotype consists of 10 experimentally determined steady-state expression patterns, considered here as the functional constraints on the network. The space of genetic networks satisfying these constraints is sometimes referred to as the neutral or genotype network. We sample this space using Markov Chain Monte Carlo which allows us to exhibit how the functional (phenotypic) constraints shape the gene network structure. We find that this shaping is strongest for the edge (interaction) usage, with effects that are functionally interpretable. In contrast, higher order features such as degree assortativity and network motifs are hardly shaped by the phenotypic constraints.

Introduction

The large regulatory networks reconstructed for *E. coli* and *S. cerevisiae* have revealed special structural features including (1) broad distributions of out-degrees, the out-degree being the number of interactions or *edges* outgoing from a given gene; and (2) the presence of *motifs*, that is repeated occurrences of small sub-graphs, the most emblematic ones being feed-forward loops[1,2,3,4,5,6]. Such genetic interaction networks provide a static genome-wide picture but tell us little about the way regulation works as a dynamical machinery. To understand better regulatory mechanisms, it is necessary to tackle time dependence of gene expression levels. This task has been undertaken on a number of small regulatory sub-systems controlling developmental processes[7,8,9], apoptosis[10], cell cycling[11,12,13] and circadian oscillations[14]. The inference of the regulatory rules in this kind of system requires detailed information on gene expression dynamics, in particular when perturbations are present (RNA interference, mutants etc.). Properties unveiled within the genome-scale networks such as the degree distributions are not easily addressed here because of the very small size of the networks. Furthermore, it is difficult to extract any regulatory principles because each system is a special case. Nevertheless, it is clear that the dynamical properties of these small networks must shape to some extent their wiring. Revealing such an effect requires more than a few network reconstructions. Here we propose to go beyond individual networks by considering *in silico* all wirings that satisfy the constraint of reproducing observed gene expression patterns. By comparing such networks to the case where the constraint is not imposed, we can see how network function shapes network structure.

To implement this program, we focus on the regulatory network controlling early flower organ specification (FOS) in the plant *Arabidopsis*. Its inferred gene regulatory network (GRN) consists of 15 genes, nearly all of which code for transcription factors. In this system as well as in many others controlling organism development, the gene expression levels of the different cell types are typically approximated as being either on or off. Based on such a Boolean framework, the group of Alvarez-Buylla provided[9,15,16,17] one of the most detailed reconstructions of a dynamical GRN (*cf.* Fig.1-A). In their model, for each gene a Boolean input to output function specifies that gene's expression level in terms of its *inputs* from the other genes. Clearly this is a crude representation of reality since it ignores the distinction between RNA and protein expression levels, pre/post transcriptional and translational mechanisms of regulation, etc. but such limitations are commonplace in GRN modelling. The Alvarez-Buylla GRN allows for 10 different steady states (Fig.1-

B) of the dynamics, associated with 10 different cell types in the early developmental stages of the *Arabidopsis* flower organs. For our study, we shall use the same Boolean framework and shall take those 10 steady states to be the *phenotypic viability*



B

	EMF1	LFY	AP2	WUS	AG	TFL1	PI	SEP	AP3	FUL	FT	AP1	*LUG*	*CLF*	*UFO*
Inflorescence I1	1	0	0	0	0	1	0	0	0	0	0	0	1	1	0
Inflorescence I2	1	0	0	0	0	1	0	0	0	0	0	0	1	1	1
Inflorescence I3	1	0	0	1	0	1	0	0	0	0	0	0	1	1	0
Inflorescence I4	1	0	0	1	0	1	0	0	0	0	0	0	1	1	1
Sepal	0	1	1	0	0	0	0	1	0	0	1	1	1	1	0
Petal p1	0	1	1	0	0	0	1	1	1	0	1	1	1	1	1
Petal p2	0	1	1	0	0	0	1	1	1	0	1	1	1	1	0
Stamen st1	0	1	1	0	1	0	1	1	1	1	1	0	1	1	1
Stamen st2	0	1	1	0	1	0	1	1	1	1	1	0	1	1	0
Carpel	0	1	1	0	1	0	1	1	0	1	1	0	1	1	0

Fig.1 The gene regulatory network constructed in the Avarez-Buylla Boolean modelling framework[9]. (A) represents the interaction graph between the 15 genes present in the GRN. The 3 external genes, having no input, are called *leaves* (orange) whereas the 12 other genes, forced to have at least one input, are *non-leaf* nodes (grey). (B) The 10 steady-states corresponding to the different organs.

constraint on our *in silico* GRNs. To test how network function shapes network structure, we consider all possible wirings and all possible Boolean rules for each gene, subject or not to the phenotypic constraint. The corresponding spaces or *ensembles* of GRNs have to be sampled rather than enumerated because of their huge size, and we do so by Markov Chain Monte Carlo[18]. In our main ensemble, referred to as *CP*, the 10 steady state constraints or *phenotypes* are imposed. In contrast, in our ensemble referred to as *C*, these constraints are not imposed. By comparing the structural features of networks in these two ensembles, we determine how the FOS function shapes the network wiring properties. We also repeat the comparison using the ensembles *CD* and *CDP* obtained from *C* and *CP* by restricting the in- and out-degrees of each gene to be those of the Alvarez-Buylla network. One of the strongest effects found in these comparisons concerns the frequency with which an interaction is present. Hereafter, we refer to these frequencies as the *edge usage* because each interaction is an edge in the graph representation of the network. The interpretation of a strong edge usage effect is that the imposition of the 10 steady states on GRNs forces certain interactions to be frequently or always present. Interestingly, network structures at higher levels, that is including more genes or edges, such as network motifs, are hardly affected by functional constraints: the few effects we do see can be understood as consequences of the edge usage structuring.

Results

The phenotypic viability constraint is severe

In the network reconstructed by Alvarez-Buylla *et al.*[9,15,16,17] shown in Figure 1-A, there are 15 genes. Three of these have no inputs, corresponding to what are called *leaves* in graph theory terminology. (We shall use the terms *node* and *gene* interchangeably in what follows.) In reality these genes must have some inputs but they probably come from outside of the list of 15 master genes included in the model. For our framework, we maintain this *leaf* status for the same 3 genes, and in fact we also maintain their property of having a single target (out-degree of 1). The other 12 genes are taken to be *non-leaf* nodes; specifically, we force them to have at least one input from one of the other genes (we thus forbid having only a self-input). The set of all networks incorporating these features is what we call the *ensemble C*. Lastly, we shall fix the number of interactions in the network. Based on the inferred network of Alvarez-Buylla *et al.*, the number of interactions will generally be 46, but all values between 12 and 147 can be considered. The size of *C* is quite astronomical; in particular, for the biological case in which the total number of edges is 46, we find (using *Mathematica* as described in the Supplementary Information) that *C* contains 1.4 10^{40} networks.

Next, we consider the ensemble *CP* obtained from *C* by further imposing what we call the *phenotypic viability* constraint: the 10 expression patterns of the FOS model must be steady states of the GRN dynamics. By construction, *CP* is contained in *C*. We find that it too is very large; for instance, for a total of 46 edges, *CP* contains 5.2×10^{37} elements. Not surprisingly, its size is substantially smaller than that of *C*. Indeed, each steady state leads to a non trivial constraint, so in practice the constraint of phenotypic viability is quite severe: for 46 edges, *CP* represents only a fraction 0.0036 of *C*, that is less than 0.4%. The sizes of these sets can be computed for any number of edges *E*. The associated numbers are displayed on a log scale in Fig.S1. From these numbers, we have calculated what we call the *CP/C* ratio, that is the fraction of networks in *C* that also belong to *CP*. We find that below *E*=16 the ratio is 0, that is one *cannot* satisfy the phenotypic viability constraint with any network if it has 15 or

less edges. The *CP/C* ratio is still small for *E*=46, but then rises rapidly. Finally one enters a regime that holds until *E* approaches its maximal value of 147, regime in which the *CP/C* ratio is very close to being linear in the number of edges, ending at *E*=147

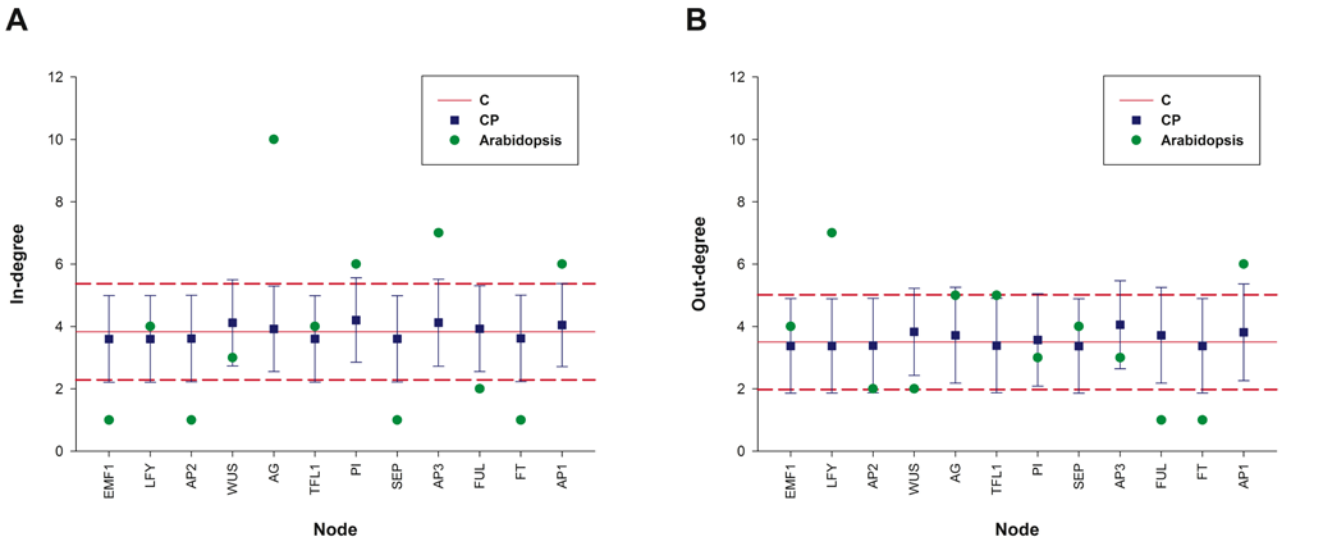


Fig.2 Comparison of each gene's in- and out-degree in the *C* and *CP* ensembles with the *Arabidopsis* reference network. (A) Ensemble mean and standard deviation of the in-degree for each non-leaf node. (B) Ensemble mean and standard deviation of the out-degree for each non-leaf node.

where *CP*=*C*. These results are displayed graphically in Fig.S2. Roughly, the condition of satisfying the 10 steady states becomes an easy to satisfy constraint beyond a number of edges (that is interactions) *E*=100.

In the case of *CD* and *CDP*, our mathematical tools are inefficient and so the exact sizes of these sets cannot be determined in practice. Nevertheless, the ratio *CDP*/*CD* can be *estimated* from the MCMC sampling of *CD* by simply computing the probability that a network chosen at random in *CD* will be phenotypically viable. With this approach, for *E*=46 interactions which will be the value imposed in most of our work, we find that the *CDP/CD* ratio is *0.13%*. From these results we conclude that whether or not one imposes the in- and out-degrees at each node, the phenotypic viability constraint is severe.

Degree distributions are affected by phenotypic viability

The in-degree (respectively out-degree) of a node is the number of interactions (edges in graph terminology) that are incoming to it (respectively outgoing from it) In our framework, the leaf nodes have in-degree of 0 and out-degree of 1, while the non-leaf nodes have variable in- and out-degrees. We will thus focus here on the non-leaf nodes only. In the *C* ensemble, phenotypes are ignored, making all non-leaf genes equivalent. Thus the ensemble mean in- and out-degree for each gene can be computed straightforwardly. Specifically, using *E*=46, the two means are 46/12=3.83 and 43/12=3.58. However the *distribution* of in- and out-degrees is non-trivial because the self-edges have a special status. In practice, the in-degree distribution is close to a shifted Poisson having one (obligatory) edge plus a number of *extra* ones that are distributed according to a Poisson law of mean 34/12=2.83. The out-degree has no particular constraint so its distribution is very close to a Poisson law of mean 43/12=3.58.

In the case of *CP*, the non-leaf nodes are no longer equivalent because of the phenotypic constraint. Thus the mean and variance of a gene's in- and out-degree will vary from gene to gene. By comparing to what is obtained without the phenotypic constraint, we find that the out-degree variances in *CP* and *C* are very similar while the in-degree variances are about 10% lower in *CP* than in *C*. In Fig.2 we display the means and standard deviations for each gene; one sees specificities for each gene, with modest gene to gene variations, these being on the order of 10 to 20%. That figure also displays the degrees for the *Arabidopsis* network[17] which we refer to as the *reference* network. Very clearly, this reference network has degrees that fall outside of the standard deviations produced in *CP* so it is fair to say that it is atypical for that ensemble. This result also follows by comparing the degree distribution of the *Arabidopsis* network and that of networks in *CP* as displayed in Fig.S3. Such a difference could be due to insufficient realism of the modelling framework as mentioned in the introduction, to uncertainties in the reference network itself[17], or to evolutionary forces outside of our scenario such as genome duplications that are known[19] to shape GRNs. No matter what, to come closer to the biologically observed properties it is appropriate to apply the standard approach[20,2] in which one fixes the in- and out-degree of each gene to the value arising in the reference network. Imposing this constraint on networks of *C* leads to the networks in the ensemble *CD*. Superposing the phenotypic viability constraints on *CD* produces the *CDP* ensemble. Comparing the networks of these two ensembles allows us to isolate the consequences of the phenotypic constraint in the same spirit as when going from *C* to *CP* while keeping structural features that follow solely from the values of the degrees.

Phenotypic viability drives specific edge usage

When imposing the phenotypic constraint, a gene must use its input signals to provide the proper output expression level for each of the 10 steady states. Since each gene has its own gene

expression profile, the choice of input edges can be important. We thus examine which edges are used and at what frequency for each gene. This information is represented via a heat map in Fig.S4-A,B for the networks in the *C* and *CP* ensemble. For the

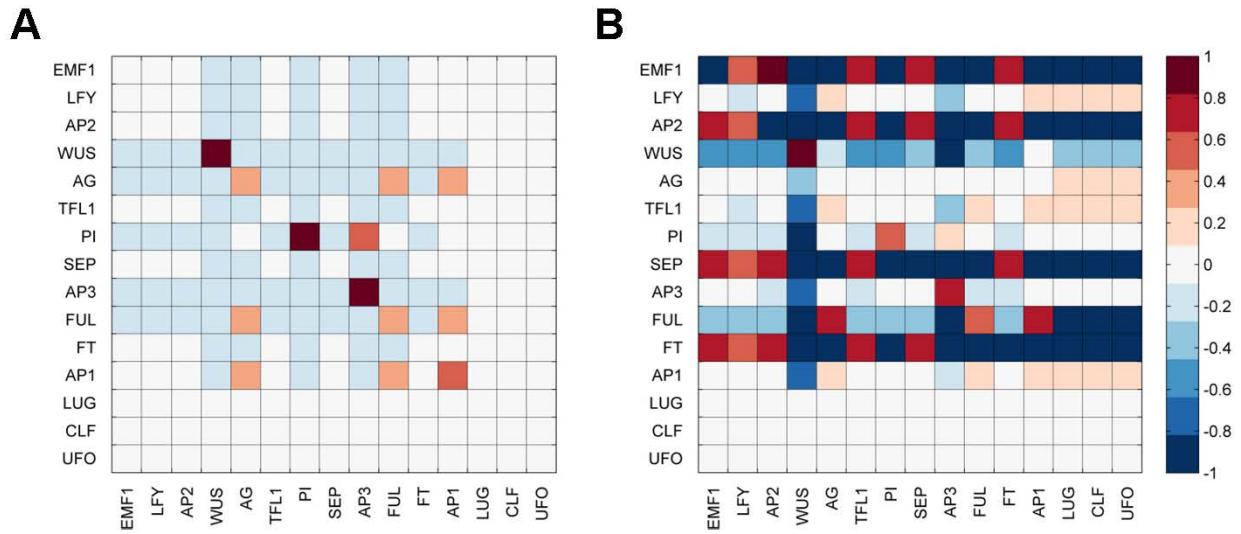


Fig.3 Change in edge usage when imposing the phenotypic constraint. Each interaction (from column number to line number) is realized at a frequency that depends on the ensemble. The natural logarithms of ratios of frequencies are shown as a heat map. (A) Log ratio of frequencies when going from *C* to *CP*. (B) Log ratio of frequencies when going from *CD* to *CDP*.

three leaf genes, no inputs are allowed, and since they each have a single output, the other genes receive their signal with an average frequency of 1/12; interestingly, in both the *C* and *CP* heat maps there are no visible deviations from this average.

Consider now the edge usage from non-leaf to non-leaf nodes. In the *C* ensemble, all non-leaf nodes are equivalent so individual genes *i* receive inputs from other genes *j* with no preferences amongst these as long as *j* is different from *i*. The result is a uniform heat map for the off diagonal elements. On the diagonal, the frequency is a bit lower because self-interactions do not count for the obligatory input to each non-leaf node. Now, for networks in *CP*, the previous section (*cf.* Fig.2) showed that the mean in- (respectively out-) degree varies slightly from gene to gene. As a consequence, the average frequency in a row (respectively column) of the *CP* heat map will depend on the gene. Interestingly, these differences in means are modest, yet, as shown in Fig.S4-B, the *CP* heat map is quite heterogeneous. First of all, the frequency of the self-edges of genes WUS and AP3 is 100%. That this has to be the case can be seen from the steady states: the expression of WUS is the only signal differentiating the steady states Inflorescence1 and Inflorescence3, while the expression of AP3 is the only signal differentiating the steady states Stamen2 and Carpel. Secondly, only two genes differentiate the steady states Sepal and Petal2: these are PI and AP3. If the gene PI has a self-edge, no other input is preferred, but if it doesn't, it must have AP3 as input. These restrictions lead to a high frequency of the edges PI → PI and AP3 → PI, as seen in the heat map. Lastly, we see in Fig.S4-B a clear pattern whereby AG, FUL and AP1 preferentially connect to one another. This pattern can be explained by observing that these 3 genes are the only ones distinguishing Petal1 from Stamen1 and Petal2 from Stamen2; they thus must receive inputs from one another or have self-interactions, leading precisely to the symmetric pattern seen in the *CP* heat map. Overall, the conclusion is that every heat map feature arising when going from *C* to *CP* can be traced to a property of the 10 expression profiles imposed for phenotypic viability. The corresponding changes of edge usage frequencies are illustrated in Fig.4-A, again via a heat map, but here based on the natural logarithms of the ratios of frequencies.

One could compare the heat map of *CP* to the actual connections in the reference network. However as seen in the previous section, the reference network has in- and out-degrees that are not typical of *CP* and so a different approach is necessary. Instead, and to refine the analysis, we will use the ensembles *CD* and *CDP* that force the number of in and out interactions for each gene to be the same as in the reference network. The heat maps for these two ensembles are shown in Fig.S4-C,D. We saw when going from *C* to *CP* that the genes WUS and AP3 necessarily had self-edges; this of course is also true and for the same reasons when going from *CD* to *CDP*. Furthermore, it turns out that the main increases in frequencies when going from *C* to *CP* simply carry over to those when going from *CD* to *CDP* (*cf.* Fig.4-A and Fig.4-B). Other differences arise, the largest one being that the frequency of presence of the LFY → SEP interaction increases when going from *CD* to *CDP* but not when going from *C* and *CP*. A plausible explanation for this fact is that the in-degree of SEP is 1, rendering it rather sensitive to the phenotypic constraint in *CDP* but not in *CP* where the individual degrees are *not* constrained.

Finally, one can ask whether the reference network is an out-lier for the *CDP* ensemble based on its edge usage. Looking at the inputs of each gene on its own, there are no cases of edge usage that seem highly unlikely for *CDP*. To quantify this, we have performed a test of the hypothesis H_0 that the reference network

belongs to the ensemble *CDP* using as summary statistic the likelihood *L* of the interactions used. For any network (actually

Table 1 Comparison of 4 directed assortativity coefficients in the Arabidopsis network with ensembles *C*, *CP*, *CD* and *CDP*. The table lists the mean and standard deviation of the 4 directed assortativity coefficients in each ensemble. The table gives the Z-scores for the 4 assortativity coefficients in the Arabidopsis network when benchmarked against the ensembles *C*, *CP*, *CD* and *CDP*.

Assortativity coefficient	Arabidopsis network	Z-score w.r.t. *C*	Z-score w.r.t. *CP*	Z-score w.r.t. *CD*	Z-score w.r.t. *CDP*	*C*		*CP*		*CD*		*CDP*	
						Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
(out,in)	-0,248	-1,298	-1,312	-0,164	0,070	-0,082	0,128	-0,084	0,125	-0,233	0,091	-0,254	0,084
(out,out)	-0,115	-0,878	-1,175	-0,361	-0,484	-0,001	0,129	0,039	0,131	-0,074	0,114	-0,063	0,107
(in,out)	-0,113	-0,817	-0,741	-0,672	-0,792	-0,004	0,133	-0,014	0,134	-0,029	0,124	-0,026	0,110
(in,in)	-0,048	-0,172	-0,400	0,583	-0,108	-0,025	0,132	0,005	0,131	-0,107	0,101	-0,039	0,083

an directed graph) *G*, *L(G)* is defined as the product over all ordered pairs (i,j) of the factor p(i←j) if the interaction i←j is present in *G* and (1-p(i←j)) if it is not, where p(i←j) is the probability of presence of that interaction as given by the heat map.

Thus we have:

$$L(G) = \prod_{\{i \leftarrow j\} \in G} p(i \leftarrow j) \prod_{\{i \leftarrow j\} \notin G} (1 - p(i \leftarrow j))$$

Within *CDP*, we compute the distribution of *L* and then determine a *p*-value for H_0 based on *L(G*)*, the value of *L* for the *Arabidopsis* reference network. We find a *p*-value of 0.19 which does not allow us to reject H_0.

Assortativity is hardly affected by phenotypic viability

Assortativity[21] in networks can be defined for any feature but in practice it is mostly used for the degree of nodes. A positive (degree) assortativity means that the degrees of neighbouring nodes are correlated positively. Investigations of numerous biological networks show that assortativity tends to be negative, *i.e.*, high degree nodes tend to connect less to other high degree nodes than in random networks. In a directed network, it is appropriate to distinguish the in- and out-degrees, leading one rather naturally to four correlation coefficients[22] for in-in, in-out, out-in and out-out degrees. We have computed these four measures of assortativity in the ensembles *C*, *CP*, *CD*, and *CDP*, and also in the reference network; the results are given in Table 1. For the four ensembles, we provide the mean and standard deviation of the assortativities, while for the reference network we provide the values and the corresponding Z-scores. The Z-score is a measure of the deviation from the mean, defined as $(x - \mu)/\sigma$ where x is the observed value while μ and σ are the mean and standard deviation in the ensemble. An absolute Z-score greater than 1.96 corresponds to an event that is outside the 95% confidence interval and thus can be considered to be an outlier.

Comparing the ensembles *C* and *CP*, we see that the phenotypic viability constraint leads to only small changes in the different assortativities (*cf.* Table 1). Nevertheless, these changes are real so phenotypic viability does affect assortativity. (Noting that each value in Table 1 is an average over 10^5 networks, the standard error on each means is of the order of one percent of the standard deviation in Table 1.) Now, using these ensembles to benchmark the *Arabidopsis* reference network, one finds that three of the eight Z-scores have an absolute value larger than 1 and that they are all negative, suggesting that the reference network is nearly an outlier for these two benchmark ensembles. Consider now the ensembles *CD* and *CDP*. Their assortativities are definitely lower than the ones of *C* and *CP*, and as expected both go in the direction of the assortivity of the reference network. Furthermore, the Z-scores of the reference network when using either *CD* or *CDP* are all less than 1 in absolute value and now take on both positive and negative values. The reference network thus seems typical for these two ensembles.

The main points brought out by these measurements are (1) phenotypic viability has negligible effects on assortativities, be-it for *C* → *CP* or *CD* → *CDP*; (2) imposing the node (gene) degrees to their values in the reference network leads to assortativities close to the ones found in the *Arabidopsis* network.

Network motifs are insensitive to phenotypic viability

In contrast to degrees that are defined for individual nodes, or edge usage and assortativity that are defined from pairs of nodes, network motifs[1,2] are defined from potentially many nodes. A motif can be thought of as a small graph with a number of nodes and edges connecting them according to a given *pattern*. For instance, a simple pattern of non-directed connections using four nodes is the square where each node connects to exactly two other nodes. Here we shall take into account the fact that the edges in our networks are directed. When defining a motif, the labels of the nodes are omitted so that only the pattern of (directed) connections matters. At a more formal graph theoretical level, motifs are subgraph topologies. From a biological perspective, if a particular motif is *over-represented*, that is arises in a network far more often than might be expected at random, then one is tempted to think that its presence is not coincidental and results from a selection mechanism.

Since a subgraph can contain smaller subgraphs, the frequencies of small motifs will influence the frequencies of larger ones just like degree distributions can influence assortativities. Let us thus begin with the simplest possible motif, a node with a self-edge, corresponding to a gene with a self-interaction. When going from *C* to *CP*, or from *CD* to *CDP*, we saw that self-interactions were obligatory for the genes WUS and AP3. We may then expect that the number of self-interactions will increase by about 2 units when enforcing the phenotypic viability constraint. As shown in Table S1, this trend occurs for both *C*→*CP* and *CD*→*CDP*. We also saw that the other major effect of phenotypic viability on the heat maps was to increase the frequencies of interactions between

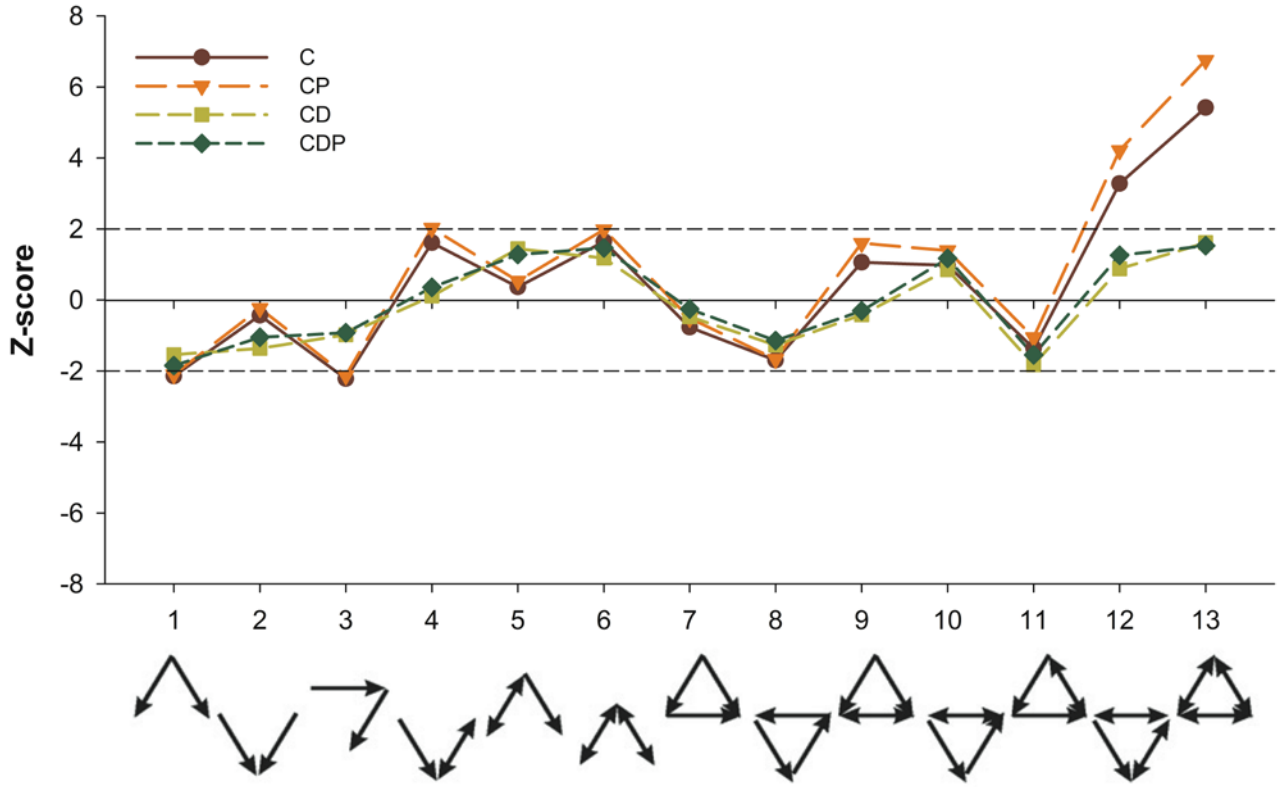


Fig.4 Motif usage in the reference network benchmarked against what arises in the four ensembles, *C*, *CP*, *CD*, and *CDP*

the three genes, AG, FUL and AP1. One may then expect phenotypic viability to increase the occurrence of the two-node motif corresponding to a mutual interaction between two genes. From Table S1, we see that the opposite happens: this motif *decreases* for both *C→CP* and *CD→CDP*. To understand this paradoxical result, recall that lower order motifs affect higher order ones. Here we know that the networks in *CP* (respectively *CDP*) have a few more self-interactions than those in *C* (respectively *CD*). Since the total number of interactions is fixed, that leaves fewer interactions to form mutual interactions, thereby lowering the expected frequency of this two-node motif. We can put this on a quantitative footing as follows. In the ensemble *C*, the distribution of the out-degree of the 12 genes that are not leaf nodes is very close to a Poisson law of mean 43/12. Taking these degrees to be independent which is an excellent approximation because only their total number is constrained, the average number of times the mutual interaction motif should appear is approximately $12\times11\times(43/12)^2/(2\times12\times12)=5.88$. This is extremely close to the actual value measured in the ensemble *C*, namely 5.81 (*cf.* Table S1). The same reasoning can now be applied to *CP* simply by taking into account the fact that some edges have to be reserved for the self-interactions. Looking at the heat map for *CP*, we see that there are typically three additional self interactions compared to what happens in *C*. Thus we will assume that there are 43-3=40 interactions available for contributing to the mutual interaction motif. Repeating the arithmetic, the average number of times the mutual interaction motif should appear is $12\times11\times(40/12)^2/(2\times12\times12)=5.09$. This is a bit lower than 5.25, the actual value measured in the ensemble (*cf.* Table S1). Thus phenotypic viability leads to a slight *enhancement* of the mutual interaction motif when using the proper background expectation. Nevertheless, it is important to keep in mind that this effect is both subtle and small.

Going on to motifs with three nodes, there are a total of 13 possible connected wirings when taking into account edge directions. Again Table S1 gives the mean and standard deviation of the occurrences of these motifs in all four ensembles. Comparing *C* and *CP*, we see that phenotypic viability reduces slightly the frequencies of all 3-node motifs, but by an amount that is only a fraction of the standard deviation, so just as for the two-node motif, the effects are small. Note that to a large extent, the decrease in the motif frequencies follows from the argument already given above: there are fewer edges able to participate in motif formation in *CP* than in *C* because of the increased number of self-interactions imposed by phenotypic viability. In the case of *CDP vs. CD*, 10 out the 13 motifs follow this rule of decreasing frequency. The same is true for the clustering coefficient[23] as shown in Table S2. The clustering coefficient measures the frequency of triangles ignoring the direction of the edges, and we find that when enforcing phenotypic viability, the frequency of the triangle motif decreases. Thus for all ensembles, the constraint of phenotypic viability seems to generate no more structuring of 3-node motifs than expected given the changes in

frequencies of the lower order motifs.

Now we will compare the number of occurrences of motifs in the *Arabidopsis* reference network to what is expected in each ensemble. Table S1 provides these numbers as well as the associated Z-scores in each case. Let us start with the single node motif: it is *under-represented* when comparing to the phenotypically viable ensembles, suggesting that there may be an artefact of our modelling that encourages self-interactions. This is quite plausible since self-interactions provide no contextual information and thus do not actually guide the system towards a good expression profile, whereas it is clearly advantageous within our modelling where we simply impose 10 steady states. Indeed, having a self-interaction for a gene guarantees that it will satisfy the steady-state constraints. Moving on to the two-node motif, we see that *Arabidopsis* has an extremely strong *over-representation* of mutual interactions. The presence of such interactions is standard lore in developmental systems: mutually inhibitory interactions allow for sharp boundaries *between* domains while mutually activating interactions allow for stable expression patterns *within* domains. However our ensembles do not exhibit as many mutual interactions as the *Arabidopsis* reference network. Just as for the case of the degrees, this *discrepancy* may be due to limitations of our modelling framework or may indicate that other factors (beyond the phenotypic constraint) affect network structure.

Finally, consider the 3-node motifs (Table S1 and Fig.4). A significant number of their Z-scores have an absolute value larger than 1.96, making *Arabidopsis* an outlier: 4 for *C*, 6 for *CP*, but 0 for both *CD* and *CDP*. This shows that it is not the phenotypic viability constraint that is responsible for *Arabidopsis* being an outlier in *CP* but rather its degree sequence. This conclusion holds also for the triangle motif (*cf.* Table S2).

Role of phenotypic robustness

So far we have considered that the phenotypic viability was the only feature to impose on GRNs to test our scenario. But even if selection acts only on phenotypic viability, as a side effect it is known that it will enhance mutational robustness in steady-state populations. We have thus investigated the consequences of mutational robustness in our GRNs on the network structural features. Since mutations break interactions, we define the mutational robustness of a GRN as the probability that it still satisfies phenotypic viability after a random edge deletion. Deleting an edge from *j* to *i* means that the Boolean value of gene *j* as seen by *i* is 0, from which we then test the phenotypic viability. Fig.S5 gives the mean robustness as a function of the number of edges for GRNs in the ensemble *CP*, showing that a maximum arises not far from 46 edges (the number in the reference network), Also shown in that figure is the average fraction of steady states that are unaffected by deleting a random edge. We then inquired whether level of robustness correlated with structural features. To do so, we divided our 10^5 networks of *CP* (respectively *CDP*) into four quartiles so that Q1 has the lowest robustness and Q4 the highest. For *CP* and *CDP* separately, these two subsets were then analysed for structural features as shown in Table S3 and S4. For the assortativities, we see that robustness changes their averages but by an amount much smaller than the width of the distributions. For the motifs, the main effect of robustness is to increase the frequency of self edges. It does not help reduce the discrepancy between our ensembles and the reference network, in particular regarding the over-representation of the mutual interaction motif.

Discussion and conclusions

The 15 master genes controlling flower organ specification (FOS) in *Arabidopsis* allow for 10 different expression profiles. By integrating results from numerous articles on physical interactions as well as on phenotypes of mutants, the group of Alvarez-Buylla was able to propose a putative interaction network with 46 edges which we call the reference network, along with Boolean rules reproducing the FOS expression profiles[9,15,16,17]. The 10 steady states of this network can be considered as part of the *function* of this genetic system. We have tackled the problem of how such a function can be implemented by more general networks and what common network structural features will thereby emerge. To quantify this *network function shapes network structure* scenario, we considered ensembles of gene regulatory networks (GRNs) using Boolean control rules and examined which structural features appeared as a result of imposing the FOS function. If a Boolean network implements the 10 FOS expression profiles as steady states, we say that it is *phenotypically viable*. In our first pair of ensembles, *C* and *CP*, *C* mainly constrains the number of genes and interactions while *CP* is the subset of phenotypically viable networks in *C*. In our second pair of ensembles, *CD* and *CDP*, *CD* is the restriction of *C* to networks having the same in- and out-degree sequence as the reference network, while *CDP* is obtained by imposing in addition phenotypic viability.

We first set out to count the number of networks before and after imposing the FOS functional constraint (Fig.1-B). For 46 interactions as in the reference network, *CP* contains more than 10^{37} networks so there is an astronomical number of ways to satisfy the phenotypic viability constraint. Nevertheless, this number represents less than a fraction 0.004 of all networks in *C*. Taking a network at random, the chance that it is phenotypically viable is tiny, and thus the phenotypic viability constraint is *severe*. The severity is worse when the number of interactions allowed in the networks is decreased, and in fact for less than 16 interactions it is simply not possible to reproduce the constraints of having all 10 steady states.

We then asked how the phenotypic viability constraint shapes structural features to test the *network function shapes network structure* scenario. The strategy used compares properties of networks in *C* and *CP* or of networks in *CD* and *CDP*. Given the huge sizes of these ensembles, we used Markov Chain Monte Carlo[18] to sample them. Thanks to this highly versatile computational tool, we generated 10^5 networks in each ensemble, allowing precise quantifications of the associated network structural properties. Starting with the simplest structural features, namely the in- and out-degrees, we found that imposing phenotypic viability had only very small effects: the distribution of degrees stayed close to Poisson, while the mean in- and out-degrees merely varied by about 5% from gene to gene (Fig.2).

This result can be qualitatively understood from the fact that the mean degree is high (non-leaf nodes have an average of 3.8 inputs): the more inputs a gene has, the more steady-state constraints it can satisfy. Imposing phenotypic viability will mainly affect the frequencies of having just one or two inputs, changing these frequencies from low to very low values. In contrast, imposing phenotypic viability had a very clear effect on the edge usage (Fig.S4): (1) self-interactions for both WUS and AP3 became obligatory, (2) PI had to have a self-interaction or an input from AP3, and (3) AG, FUL and AP1 preferentially connected to one-another. All these effects were interpretable using the 10 steady state expression patterns. In a more general context, one may expect such edge usage shaping to be all the more visible that there are many constraints (here steady states) to satisfy and few edges to work with.

We also studied higher order structural features of networks in our ensembles: assortativity, motifs and clustering (which is associated with the triangle motif ignoring edge direction). In all these cases, the changes due to imposing phenotypic viability were generally understood through the shaping of the lower order features, here node degree and edge usage; for instance the self-interactions drove down slightly all higher order motifs. Our conclusion is thus that phenotypic viability here has hardly any direct consequences on higher order structures (*cf.* Fig.4 and all Tables), a result that had also been seen in another *in silico* system[24]. In contrast, there were quite large changes when going from *C* to *CD* or from *CP* to *CDP* simply because structural features depend significantly on the in- and out-degree sequence. In a similar vein, we also asked whether the phenotypic *robustness* of networks shapes their structural features. The answer is yes, although in the system studied here the effects are small (*cf.* Tables S3 and S4). Perhaps the main consequence of robustness is to act on the *number* of interactions; as shown in Fig.S5, robustness is maximized when the number of interactions is close to that found in the reference network.

Finally, the *network function shapes network structure* scenario was used to compare the *Arabidopsis* network proposed by Alvarez-Buylla *et al*. to the networks in our ensembles. Not surprisingly we recovered some of the features of the edge usage that are interpretable in terms of the steady-state constraints, just as when going from *C* to *CP* or *CD* to *CDP* (*cf.* Fig.3). Beyond that, it transpired that for the in- and out-degree sequences, the *Arabidopsis* network was a clear outlier, having many more genes than expected with very low or very high degree (*cf.* Fig.2 and Fig.S3). Such a property may be due to limitations of the Boolean GRN modelling framework that ignores the detailed molecular nature of interactions. In particular, all interactions are counted in the same way, be-they transcriptional or translational, thereby probably biasing the use of interactions in our ensembles. It is also possible that evolutionary history has played a role in addition to the functional constraints we focused on in this work. In particular, there have been studies modelling the effects of genome duplications on regulatory motifs[19]; in the *Arabidopsis* network some of the genes are believed to be distant paralogs, a property that would lead to an enhancement of particular motifs such as the 2-node mutual interaction. In spite of these possible limitations, it should be clear that this system has exhibited certain network structures that are consequences of network function, even if it is mainly at the level of edge usage rather than the more appealing level of motif over-representation.

Model and Methods

Boolean Gene Regulatory Networks

Boolean gene regulatory networks[25,26,27] are simplified models of gene regulation that take the state of each gene to be either on or off. The associated expression levels are typically represented by the binary values 1 and 0. In a GRN there are nodes representing genes as well as directed edges representing genetic interactions that specify which genes serve as input to which. For a given gene, the input to output relation is necessarily Boolean, and can be represented by a truth table in which the output (0 or 1) is specified for all possible 2^k input patterns if there are k inputs. The set of interactions and the Boolean functions for each gene can be considered to define the genotype of the network.

The *phenotype* of a GRN depends on the expression patterns generated by the network under its expression dynamics. These depend on the Boolean functions for each gene but also on the order in which gene expression values are updated. The group of Alvarez-Buylla[9,15,16,17] as well as many other authors use discrete time steps with synchronous updating, that is all genes are updated at the same time at each step. Given the specification of the dynamics, the genotype completely defines the phenotype so we have a genotype to phenotype map. In the present work, for the phenotype we focus only on the steady states of the dynamics, and therefore the order of the updating does not matter.

The Arabidopsis FOS GRN

By integrating results from a large number of publications, the group of Alvarez-Bullya was able to propose a list of 15 genes and 46 interactions that drive flower organ specification (FOS) in *Arabidopsis thaliana*. Of these 15 genes, 3 are *leaves*, *i.e.*, are nodes with no inputs. Thus, in the context of the model, their expression levels are externally determined, independently of the rest of the network; these leaf nodes each connect to the non-leaf nodes with one edge, and have no inputs. The 12 other genes have inputs from at least one gene other than themselves.

Going beyond a static description, Alvarez-Buylla's group furthermore constructed a discrete time dynamical Boolean GRN that has the property of reproducing the 10 known expression patterns in the different parts of the flower primordium. These patterns are provided in Fig.1. The Boolean rules for the different genes have been validated both using the *Arabidopsis thaliana* wild type and various mutants[9,15,16,17].

The Ensembles *C, CP, CD* and *CDP*

To understand the consequences of phenotypic viability (the phenotypic constraint) on the structure of gene regulatory networks, we constructed 4 ensembles of networks that are subject to different sets of constraints. For the ensemble "*C*", each leaf node has one unique output connection to a non-leaf node and no incoming edge; furthermore, non-leaf nodes must always have at least one incoming interaction and when there is

in fact only one, it may not be a self-input. These constraints are motivated by the interactions observed in the *Arabidopsis* reference network. Finally, we also take the total number of edges to be fixed; for most of the results shown, this number is set to 46, the number of edges in the reference network.

For the next ensemble, referred to as "*CP*", we begin with the *C* ensemble and impose in addition the *phenotypic* constraint, thus the "*P*" in this nomenclature. Explicitly, given a network of interactions satisfying the constraints of *C*, membership in *CP* requires that it be possible to find Boolean input-output rules such that the 10 *Arabidopsis* expression profiles are all steady states of the dynamics. If that is the case, we call the network *phenotypically viable*. Clearly, comparing *CP* to *C* directly brings out the consequences of satisfying the *Arabidopsis* phenotypic constraint.

The ensemble "*CD*" is the subset of *C* where each gene has the same *in-degree* and *out-degree* as in the reference network. *CD* corresponds in effect to the set of networks obtained by applying the edge exchange algorithm[20,2] to the *Arabidopsis* reference network.

Lastly, the ensemble "*CDP*" is the subset of *CD* consisting of all of its phenotypically viable networks. Just as for the passage from *C* to *CP*, going from *CD* to *CDP* exhibits the consequences of the *Arabidopsis* phenotypic constraint, but where possible effects of the degrees have been already been taken into account.

MCMC sampling of each ensemble

It is not possible to consider all networks in any of our ensembles because they are far too many. The next best thing to do is to sample each ensemble sufficiently to obtain its properties to whatever precision is necessary. Markov Chain Monte Carlo (MCMC) makes this possible even in the ensembles where the networks are highly constrained (for instance by the phenotypic constraint). The principle is completely general: one sequentially constructs a succession of networks N_1, N_2, … , N_n, all belonging to the ensemble under consideration. Each network is constructed from the preceeding one using a Markov Chain: a proposed change is stochastically generated and then it is either accepted or rejected according to the so called Metropolis rule[18]. In practice, the proposed change consists in removing one edge and replacing it by another. The Markov Chain must allow an exploration of the whole space, and then the Metropolis rule ensures that the sampling is asymptotically uniform[18]. In practice, the MCMC runs must be long enough to be in that asymptotic regime, and this can be checked self-consistently by measuring the auto-correlation time of the Markov chain. For the present ensembles, the auto-correlation times are short (corresponding to a few proposed changes per gene) and so the MCMC is both reliable and efficient.

Network structural features

Structural features of networks range from local properties (such as node degrees) to global properties (such as connectivity). In the present framework, the network considered is the directed graph of interactions embedded in the gene regulatory network of interest. The most local structural feature is the in-degree and out-degree of a node (specifying a gene). For an *ensemble* of networks, one can consider the distribution of these quantities at a given node; we focus on the mean and standard deviations for each gene separately. On can also study the *edge usage* in the ensemble, given by the probability of occurrence of each edge (from gene *j* to gene *i* for all ordered pairs of genes). For convenience, we represent these probabilities by a heat map as in Fig.S4.

The local clustering coefficient C_i at node *i* focuses on that node and its immediate neighbors[23]. C_i is defined as the number of edges connecting any pair of these nodes divided by the maximum number that could be present. It thus measures the overlap between the sub-graph containing these nodes and the clique in which all of these nodes are pairwise connected. Averaging C_i over all nodes of the graph gives the network's *average clustering coefficient* C. In an ensemble of networks, C will have a distribution that can be used to benchmark whether a given network is outlying or not for clustering.

A closely related index is the so called *assortativity coefficient* for degrees[21]. It quantifies the correlation of the degrees between neighboring nodes. For directed graphs, one can define 4 different measures of assortativity, obtained by calculating in-in, in-out, out-in, and out-out degree correlations over all nodes[22]. Again, in an ensemble of networks, the assortativity coefficient will have a distribution from which one can perform a test of whether a given network is outlying or not for assortativity.

Network motifs[1,2] are even less local structural features of biological networks. Each corresponds to a pattern of connected nodes such as a triangle or a square. In a more mathematical definition, motifs are sub-graphs in which nodes have no labels. In the present work, we deal with directed graphs so the motifs have directions on their edges. Most studies on network motifs involve proper graphs in which there are no self-interactions, and generally one forces the motif to form a connected sub-graph. Then there are two motifs containing two nodes and 13 containing three nodes as drawn at the bottom of Figure 4. For any network motif, one can compute the number of its occurrences in a given graph, and if desired compare to what is expected in an ensemble.

Acknowledgements

We thank Daniel Baker, James Eddy, Yulang Luo, Ilya Shmulevich, and Joseph Xu Zhou for discussions. AS acknowledges support from CNRS GDRE513. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. S. S. Shen-Orr, R. Milo, S. Mangan and U. Alon, Nature genetics, 2002, **31**, 64-68.
2. R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii and U. Alon, Science, 2002, **298**, 824-827.
3. T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J.

Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J. B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford and R. A. Young, Science, 2002, **298**, 799-804.
4. A. L. Barabasi and Z. N. Oltvai, Nat Rev Genet, 2004, **5**, 101-113.
5. M. M. Babu, N. M. Luscombe, L. Aravind, M. Gerstein and S. A. Teichmann, Current opinion in structural biology, 2004, **14**, 283-291.
6. M. M. Babu, Biochemical Society transactions, 2010, **38**, 1155-1178.
7. G. von Dassow, E. Meir, E. M. Munro and G. M. Odell, Nature, 2000, **406**, 188-192.
8. R. Albert and H. G. Othmer, Journal of theoretical biology, 2003, **223**, 1-18.
9. C. Espinosa-Soto, P. Padilla-Longoria and E. R. Alvarez-Buylla, The Plant cell, 2004, **16**, 2923-2939.
10. H. J. Huber, M. Rehm, M. Plchut, H. Dussmann and J. H. Prehn, Bioinformatics, 2007, **23**, 648-650.
11. K. C. Chen, L. Calzone, A. Csikasz-Nagy, F. R. Cross, B. Novak and J. J. Tyson, Molecular biology of the cell, 2004, **15**, 3841-3862.
12. F. Li, T. Long, Y. Lu, Q. Ouyang and C. Tang, Proceedings of the National Academy of Sciences of the United States of America, 2004, **101**, 4781-4786.
13. M. I. Davidich and S. Bornholdt, PloS one, 2008, **3**, e1672.
14. J. C. Leloup and A. Goldbeter, Proceedings of the National Academy of Sciences of the United States of America, 2003, **100**, 7051-7056.
15. A. Chaos, M. Aldana, C. Espinosa-Soto, B. G. P. de Leon, A. G. Arroyo and E. R. Alvarez-Buylla, J Plant Growth Regul, 2006, **25**, 278-289.
16. E. R. Alvarez-Buylla, A. Chaos, M. Aldana, M. Benitez, Y. Cortes-Poza, C. Espinosa-Soto, D. A. Hartasanchez, R. B. Lotto, D. Malkin, G. J. Escalera Santos and P. Padilla-Longoria, PloS one, 2008, **3**, e3626.
17. E. R. Alvarez-Buylla, M. Benitez, A. Corvera-Poire, A. Chaos Cador, S. de Folter, A. Gamboa de Buen, A. Garay-Arroyo, B. Garcia-Ponce, F. Jaimes-Miranda, R. V. Perez-Ruiz, A. Pineyro-Nelson and Y. E. Sanchez-Corrales, The Arabidopsis book / American Society of Plant Biologists, 2010, **8**, e0127.
18. Metropolis N, Rosenbluth A, Rosenbluth M, Teller A and Teller E, Journal of Chemical Physics, 1953, **21**, 1087-1092.
19. S. A. Teichmann and M. M. Babu, Nature genetics, 2004, **36**, 492-496.
20. S. Maslov and K. Sneppen, Science, 2002, **296**, 910-913.
21. M. E. Newman, Physical review letters, 2002, **89**, 208701.
22. J. G. Foster, D. V. Foster, P. Grassberger and M. Paczuski, Proceedings of the National Academy of Sciences of the United States of America, 2010, **107**, 10815-10820.
23. D. J. Watts and S. H. Strogatz, Nature, 1998, **393**, 440-442.
24. J. F. Knabe, C. L. Nehaniv and M. J. Schilstra, Bio Systems, 2008, **94**, 68-74.
25. S. A. Kauffman, The origins of order : self organization and selection in evolution, Oxford University Press, New York, 1993.
26. S. Kauffman, Nature, 1969, **224**, 177-178.
27. S. A. Kauffman, Journal of theoretical biology, 1969, **22**, 437-467.