SI Appendix

1 Leaf Removal Algorithms

In this note, we provide more information and results relative to the evaluation of feedback with leaf-removal algorithms. We considered three variants of this decimation algorithm that removes the tree-like parts of a graph, leaving the components with feedback (1). The graph has N nodes, of which M are regulated. Define a leaf as a node having only incoming links, and a "free" variable, or a root, a node having only outgoing links. At each iteration, the algorithms identify and remove a set of links and nodes from the graph, using the following prescriptions (SI Fig. 4).

1. LRa. Remove leaves and their incoming links.
2. LRb. As LRa. Additionally remove incoming links of nodes whose incoming links are all connected to roots, which are also removed.
3. LRc. As LRa. Additionally, remove all the incoming links (together with their associated nodes) of nodes whose incoming links are connected to at least one root.

Note that multiple nodes may disappear in a single move. There are two possible outcomes for the leaf removal. Reaching the free nodes, which are left as isolated points, or halting when a "core" graph that contains feedback is present. The core is composed of N_C nodes and M_C constraints. One can use these as order parameters, i.e. to measure the extent of feedback. Equivalently, one can use $\Delta_C = \frac{N_C - M_C}{N}$ or $\gamma_C = M_C/N_C$. The difference between LRa and LRb is that LRb is able to remove tree-like parts of the graph that lie upstream of a simple loop. LRc is also able to do this. On the other hand, LRc might break some of these loops if they are connected to roots (SI Fig. 4). LRc cannot break "complex" loops, or loops not connected to roots by a single link. This is related to the existence and clustering of solutions in a random Boolean model associated to the graph (1–3). SI Fig. 5 reports the histograms of the core sizes and order parameters of randomized counterparts of the *E. coli* network using the different variants of leaf removal. No significant difference can be detected between LRa and LRb, while LRc typically yields an empty core. This indicates some sort of "simplicity" in the feedback of the randomized counterparts. The LRa and LRb core histograms can be compared with the results on model graphs (1). This comparison suggests that *E. coli* randomized counterparts might be in the region at the borderline between an empty-core and a nonempty-core regime.

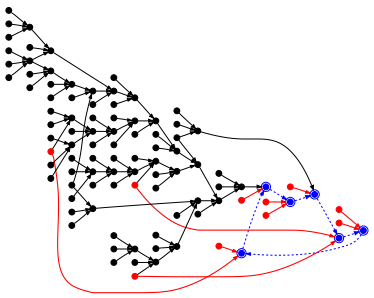

Figure 4: Examples of cores for the different leaf-removal variants, applied on the same initial graph. The cores are superimposed. The LRa core (whole figure) contains feedback loops and tree-like regions (solid edges, black) upstream of the loops. The LRb core (thick edges, red) does not contain the tree-like parts, but all the feedback is preserved. The LRc core is empty, as this algorithm is able to break simple loops connected to single free variables. The loop present in the original graph is indicated by circled nodes and dashed edges (blue)

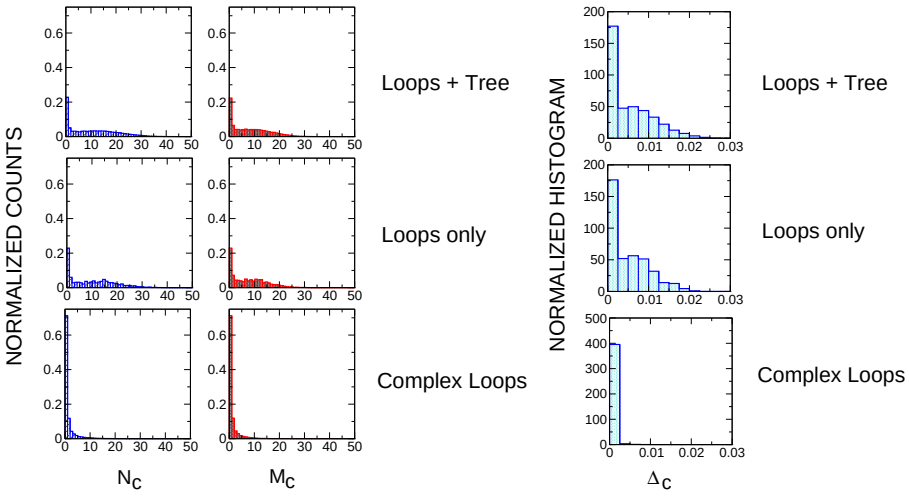

Figure 5: (a) Core size for the different variants of the leaf-removal algorithm applied to *E. coli*. Top LRa. Middle LRb. Bottom LRc. (b) Histogram of the order parameter Δ_c for different variants of the leaf-removal algorithm applied to *E. coli*. Top LRa. Middle LRb. Bottom LRc. The data refer to $1.1 \cdot 10^6$ accepted MCMC moves for randomization. The graph shows that LRa and LRb give essentially the main result, with roughly 20% of the randomized network counterparts having nonempty core. On the other hand LRc gives typically an empty core.

2 Additional Results on Gene Duplication

Here, we provide some additional results relative to the analysis of network structure in comparison with classes of likely duplicate genes. We first discuss an alternative method to define the homology classes using SUPERFAMILY (4) domain architectures, then present additional data on the partition of classes among computational layers, and the few existing links among TFs belonging to the the same homology class.

2.1 Alternative Criteria for Homology Classes

The duplication analysis in the body of the paper defines homologous proteins through structural domain assignments of the SUPERFAMILY database (4). Specifically, homologs are by definition proteins whose domain architectures are identical, respecting domain order and neglecting domain repeats. We refer to this as criterion A. This criterion is similar to those used in refs. 5 and 6. Since the definition homology is rather arbitrary, it is natural to explore different variants and verify that the results are stable. To this aim, we considered a criterion (criterion B) that allows for partial overlaps in domain architecture. Homology is defined by the existence of a subsequence of shared domains in the two domain architectures. This means that members of the same class can have acquired new domains during evolution. In both cases, the existence of a gap in different positions of the protein was considered relevant for classification. In other words, a gap was interpreted as a domain, though of unknown origin. The reason for this hypothesis is that gaps could correspond to functional domains that are not recognized by the database. We also considered the same two criteria where gaps were ignored, obtaining once again the same qualitative results (data not shown). Note that the definition of a gap is by itself an issue. The above analysis relies on the data set of ref. 6 for gap definition. We tested our results with domain architectures obtained directly from the SUPERFAMILY 1.69 database, and gap definitions of different lengths, obtaining no significant difference in the results.

The TF classes obtained with criteria A and B are reported in SI Fig. 6 . The left column contains the SUPERFAMILY domain architecture defining each class. The right column contains the list of genes belonging to that class (classes entirely contained in other classes are not depicted). The label `AR_` marks the class members that have an autoregulatory link in the transcription network. With criterion B one obtains larger classes, but also the AR population changes accordingly. SI Fig. 7 reports the AR population of the classes for the case of criterion B. In this case, more noise is present in the statistical signal. For example, the P-value for ARs emerging from duplication is lower (about 20%) if one considers the same global observables as for criterion A, g_{AR} and h_{AR} defined in the text. However, observations on single classes lead to high confidence for AR duplications. For example, the P-value for the observed AR population of the largest homology class (that has 40 members) is 1%.

CRITERION A

```
46689,_gap_       acrr AR_beti AR_uidr
_gap_,46689       adiy appy envy
52172,_gap_       arca AR_cpxr evga kdpe narl ompr AR_phob AR_torr uhpa yjdg
46785             AR_arsr AR_emrr AR_exur AR_fur glcc AR_marr AR_pdhr
                  AR_uxur yjbk
46785,69732       AR_asnc AR_lrp
46785,53850       cbl AR_cynr AR_cysb AR_dsdc AR_gcva AR_hcar AR_ilvy leuo
                  AR_lysr AR_metr AR_nac nhar AR_oxyr AR_tdca xapr
51206,46785       AR_crp AR_fnr
_gap_,46894       AR_csgd AR_rcsa
47413,53822       cytr ebgr frur AR_gals gntr AR_idnr laci AR_mali AR_purr
                  rbsr trer
_gap_,52540       dnaa rtcr
46785,_gap_       AR_exur AR_fur glcc AR_pdhr AR_uxur yjbk
48329             AR_flia rpos AR_rpoe rpoh
46785,52512       glpr AR_srlr
47729             AR_hima AR_hns
46785,55781       iclr mhpr yiaj
46689,46689       AR_mara AR_soxs
51182,46689,46689 AR_melr AR_rhas AR_rhar
46785,_gap_,53067 mlc AR_nagc
46955,_gap_       AR_soxr zntr
```

CRITERION B

```
46689,_gap_       AR_uidr AR_beti AR_ada acrr
_gap_,46689       envy appy adiy
52172,_gap_       yjdg uhpa AR_torr AR_phob ompr narl kdpe evga AR_cpxr arca
46785             yjbk yiaj xapr AR_uxur AR_tdca AR_srlr AR_pdhr AR_oxyr nhar
                  AR_nagc AR_nac mode mlc mhpr AR_metr AR_marr AR_lysr AR_lrp
                  AR_lexa leuo AR_ilvy iclr AR_hcar glpr glcc AR_gcva AR_fur
                  AR_fnr fadr AR_exur AR_emrr AR_dsdc AR_cysb AR_cynr AR_crp
                  cbl bira AR_asnc AR_arsr AR_argr
46785,69732       AR_lrp AR_asnc
52172,52540       AR_glng atoc
46785,53850       xapr AR_tdca AR_oxyr nhar AR_nac AR_metr AR_lysr leuo
                  AR_ilvy AR_hcar AR_gcva AR_dsdc AR_cysb AR_cynr cbl
51206,46785       AR_fnr AR_crp
_gap_,46894       AR_rcsa malt AR_csgd
47413,53822       trer rbsr AR_purr AR_mali laci AR_idnr gntr AR_gals frur
                  ebgr cytr
_gap_,52540       rtcr dnaa
_gap_,46785       mhpr leuo AR_emrr
46785,_gap_       yjbk AR_uxur AR_pdhr AR_nagc mlc glcc AR_fur AR_exur
48329             rpoh AR_rpoe rpos AR_flia
46785,52512       AR_srlr glpr
47729             AR_hns AR_hima
47413             trer rbsr AR_purr nadr AR_mali laci AR_idnr AR_hipb gntr
                  AR_gals frur ebgr deor cytr
46785,55781       yiaj mhpr iclr
46689,46689       AR_soxs rob AR_rhar AR_rhas AR_melr AR_mara AR_arac
51182,46689,46689 AR_rhar AR_rhas AR_melr
46785,_gap_,53067 AR_nagc mlc
52540,48283       pspf AR_glng
_gap_,53067       AR_alsk AR_nagc mlc
46955,_gap_       zntr AR_soxr
48283             AR_fis pspf AR_glng
```

Figure 6: TF homology classes defined with the two different criteria from SUPERFAMILY domain architectures. The label AR_ indicates nodes with self-links.

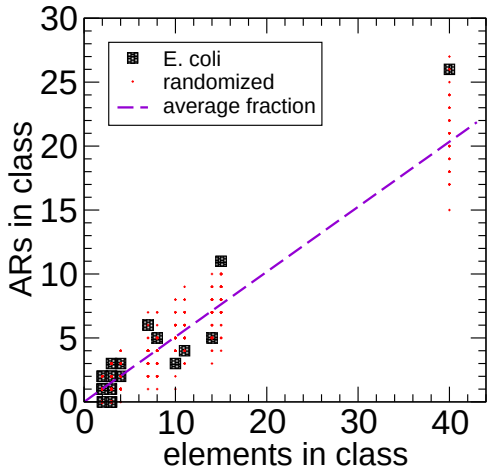

Figure 7: Population of ARs in the classes defined by criterion B. The graph correspond exactly to that reported in Fig. 2a in the body of the paper. The x axis reports the size of each class of transcription factors, while the y axis indicates the fraction of autoregulators in the class. The dashed line corresponds to the expected value computed from the total fraction of ARs. Red dots are randomized instances.

2.2 Distribution of Classes in Layers

SI Fig. 8 shows how genes of the *E. coli* network are distributed among the different classes (criterion A).

2.3 Duplicated TFs with crosstalk

SI Fig. 9, shows the 5 pairs of duplicate genes that present an intra-class link, which could be a reminiscence of an initial crosstalk. The pairs are shown together with their immediate surroundings in the network.

2.4 Function of Duplicated ARs

We find a tendency of the ARs that populate the same class to have the same function in the corresponding self-links. To quantify this, SI Fig 10 reports the histogram of the quantity $a_{AR} = (n_a - n_r)/n_{AR}$, where n_a, n_r, n_{AR} are, respectively, the number of activators, repressors, and autoregulators in the class. ARs of the same homology class tend to have the same annotated function, hence the peaks at ± 1 (particularly at -1, since most of the ARs are self-repressors). Comparing to randomizations, this observation has very high P-value for the larger classes ($< 10^{-5}$).

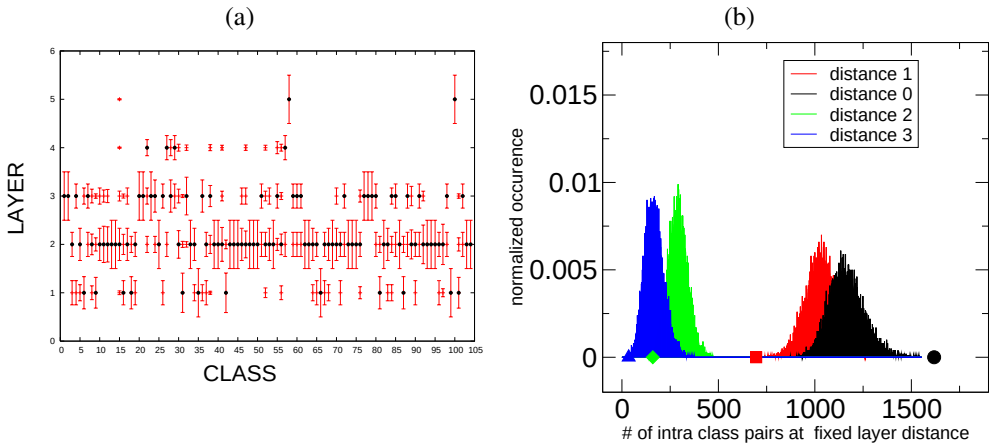

Figure 8: (a) Population of computational layers by the the homology class members defined by criterion A. The x-axis reports the homology classes, and the y-axis the computational layers. The bars indicate the fraction of the class population in a given layer. The black dots correspond to the most populated layer for that class. (b) Comparison with randomized counterparts. Each color corresponds to a different inter-class distance. The histograms represent the population of genes at a given inter-class distance in the randomized case, while the dots of the same color correspond to the observed values in *E. coli*.

3 Theoretical Considerations on Duplication

3.1 Propagation of Autoregulatory Nodes and Crosstalks

In this note, we present some simple theoretical arguments on the duplication of autoregulatory nodes. We will reason using the graph growth model introduced in the body of the paper, relative to the proliferation of ARs and intra-class links after duplication events. The model is specified as follows. Each evolutionary growth step is composed by two substeps. The first substep is gene duplication, $a \longrightarrow a + a'$ as illustrated in SI Fig. 11. In this substep, links are added in in a way that each duplicate inherits all the interactions of the original. The second substep models divergence, and corresponds to cancellation of regulatory links with different probabilities. Probabilities to keep a self-link are noted p_{Aa} and $p_{A'a'}$, while cross-regulatory links are noted $p_{A'a}$ and $p_{Aa'}$ (A is the protein produced by gene a).

In the following, we will argue that it is not automatic to propagate a finite fraction of ARs considering only duplication-cancellation processes, and that additional mechanisms are needed in the evolutionary dynamics. To this end, we initially assume that the transcriptional regulatory network evolve only by local, partial or possibly global duplication, *i.e.* at each step a fraction q of all N nodes of the graph are duplicated.

Let us first consider the (null) hypothesis whereby any regulatory link has the same probability to disappear following gene duplication, implying in particular that self-links are on the same footing as other regulatory links. This can be viewed as a totally symmetric situation for

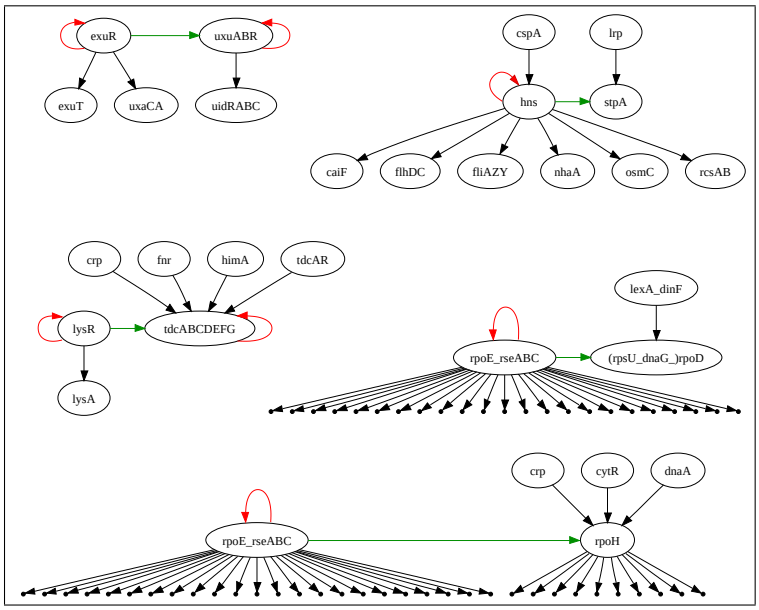

Figure 9: Local subgraphs in the *E.coli* transcription network, relative to pairs of homologous genes connected by a regulatory link. The graphs, refer to both criterion A and B; the pair lysR-tdcA is present only with criterion B. In all the five examples, at least one member of the pair has one self-regulatory link (in red). Links within classes are in green.

divergence, with $p_{Aa} = p_{A'a'} = p_{A'a} = p_{Aa'} = p$, $p \in [0,1]$. In this case, if the graph at time t has N_{AR} self-links, the subsequent step will bring $N \to (1+q)N$, and $N_{AR} \to (1+q)pN_{AR}$, where we assume for the moment that all duplicated genes are kept *irrespective* of their mode of regulation. The evolution equation for the fraction of ARs, $f_{AR} := N_{AR}/N$, is then

$$f_{AR}(t+1) = p f_{AR}(t) \quad . \tag{1}$$

The same reasoning can be extended to the more realistic "asymmetric" hypothesis where at least one self-regulatory link is preserved per duplicated pair of ARs, *i.e.* $p_{Aa} = 1$ ("old" gene) and $p_{A'a'} = p \in [0,1]$ ("new" gene), with $p_{A'a}$ and $p_{Aa'} \in [0,1]$. In this case, one finds for the fraction of ARs,

$$f_{AR}(t+1) = \frac{1+p}{2} f_{AR}(t) \tag{2}$$

The important fact implied by Eqs. 1 and 2 is that, in the hypothesis of "neutrality", the fixed point for the fraction of ARs is always zero. Biologically, the necessity for a fixed point can be argued in connection with the ability of gene duplication to drive the abundance of a certain feature on long time scales. In the hypothesis where ARs are on the same ground as non-ARs, duplication cannot, in the long run, sustain the observed nonvanishing fraction of autoregulators. Moreover, it is easy to show that the only possible fixed point remains zero in the asymmetric case, if a fraction of the *new* nodes are kept together with all the originals.

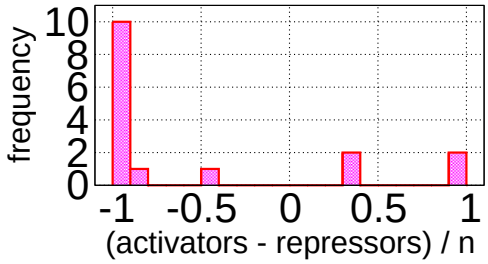

Figure 10: Histogram of the homogeneity of the function on the self-links of ARs belonging to he same homology class. Dual links are not kept into account.

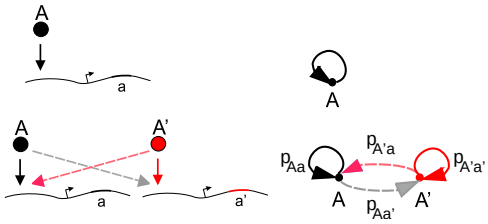

Figure 11: Sketch of the model for duplication and cancellation of links. At each time step a fraction of genes is duplicated and all the interactions among duplicates are inherited. Subsequently, divergence can erase links with probabilities p_{Aa} and $p_{A'a'}$ for self-links and $p_{A'a}$ and $p_{Aa'}$ for cross-regulatory links (A is the protein produced by gene a).

Hence, this implies that the observation on the fraction of ARs reported in the body of the paper must be due to additional mechanisms, that break the symmetry between AR and non-AR nodes, and lead to a map with a nonvanishing fixed point. Two candidate mechanisms are the following.
(i) Selective pressure for AR nodes. In other words, at each divergence substep of the dynamics, a fraction $q' > q$ of AR nodes is duplicated. With this mechanism, one obtains a finite fixed point for ARs without any rewiring. For example, in the asymmetric case, Eq. 1 becomes

$$f_{AR}(t+1) = \frac{t f_{AR}(t)}{1 + (s-1) f_{AR}(t)} \quad , \tag{3}$$

where $s := (1+q')/(1+q)$ and $t := (1+pq')/(1+q)$ play the role of selection factors. It is easy to check that, for $p > q/q'$, Eq. 3 has the nonzero fixed point

$$f_{AR}^{\text{fix}} = \frac{t-1}{s-1}$$

An analogous argument holds for the symmetric case.
(ii) A "source" term for ARs, which could come from spontaneous creation of AR links by rewiring, or horizontal transfer of ARs (together with their self-link). Suppose for example that self- and cross- (non-self) links are created spontaneously with the same probability p_L. As an

upper bound for p_L, we can take the fraction of links in the network, $p_L \simeq L/NN_{TF}$, where L is the number of links, and N_{TF}, N the number of TFs and the total number of nodes respectively. One can then write in the asymmetric duplication case,

$$f_{AR}(t+1) = \frac{1+p}{2} f_{AR}(t) + p_L \left(1 - \frac{1+p}{2} f_{AR}(t)\right) \quad , \tag{4}$$

yielding the fixed point

$$f_{AR}^{\text{fix}} = \frac{p_L}{1 - \frac{1+p}{2}(1-p_L)} \, .$$

Using the *E. coli* network data, one can work out the value of p needed to give the observed AR fraction (about $1/2$), obtaining $p \simeq (3p_L - 1)/(p_L - 1)$. This is close to $p = 0.97$.

Lastly, one can build a similar argument for the observed average number of cross- links (in number N_I) per node, within one class, $f_I := N_I/N$ (see SI Fig. 9). The evolution equations for the number of cross-links involve explicitly the fraction q of duplicated nodes, and are complicated by the combinatorics of the possible duplications. For simplicity, we restrict the discussion to the symmetric case, where self- and cross- links are kept with the same probability p. We obtain the following equation for the evolution of cross-links,

$$f_I(t+1) = p \left(\frac{2q}{1+q} f_{AR}(t) + (1+q) f_I(t) \right) \, . \tag{5}$$

Equation (5) determines the ratio of cross- to self- links. If $pq > 1 - p$, it implies that f_I does not have a fixed point, but diverges (and thus trivially dominates). Otherwise, one finds

$$\frac{f_I^{fix}}{f_{AR}^{fix}} = \frac{2pq}{(1+q)(1-p(1+q))} \, ,$$

where we supposed the presence of a fixed point f_{AR}^{fix} for the ARs.

The important point here is that ARs are a source term for intra-class cross-links (as can be seen in SI Fig. 11), while the opposite is not true. As a consequence, cross-links will typically dominate the link population of a class, unless explicitly suppressed. For $pq < 1 - p$, one can consider the ratio of ARs to the total number of links in the class $R = \frac{f_{AR}}{f_{AR}+f_I}$, as done in the text. For example, if $q = 1$ and $p < 1/2$, $R = (1-2p)/(1-p)$. This can be compared to the observed value of 91% for R, suggesting the systematic cancellation of crosstalks between duplicate ARs that we already discussed in the body of the paper. In fact, to obtain such a high value for R with our equation, one would have to use the extremely low value $p \simeq 0.08$.

In summary, we have argued that

1. A duplication dynamics where ARs are on the same status as other nodes leads to a vanishing fraction of ARs.

2. A finite fraction of ARs could be obtained either by a selective pressure for AR nodes or by an influx of ARs by rewiring or addition of AR nodes.

3. AR duplications are a source for links between nodes in the same homology class, which would typically dominate the population of links within a class if not erased by selective pressure.

4 The pair CRP-FNR

4.1 Specificity of the duplicate TF pair

All the above considerations are based on a large-scale, or "coarse-grained" view of the regulation network. It is useful to try to reconstruct these observations using more microscopic analyses. As an example, we considered the specificity of the two homologs FNR and CRP. This is an interesting pair, as the two ARs have similar consensus sequence, but lack any crosstalk, while they both regulate a very large number of targets (SI Fig. 12). We clustered their binding sites found in the regulonDB database (7) using simple concepts of information theory (see the following section) (8). Our method can account for the specificity of about 85% of the binding sites, while the others remain unexplained. Of particular interest, are four binding sites where the two TFs are reported to overlap (relative to the transcription units of acnA, focA-pflB, glpABC, hlyE). Among these four targets, we find that only two (acnA, focA-pflB) seem to have marked specificity for FNR, while the other two have specificity zero for both TFs, and thus are confirmed as overlapping by our method. Finally, we find that all of the four autoregulatory binding sites have high specificity (SI Fig. 12), which accounts for a divergent coevolution process which suppresses crosstalk. Regarding the mechanism for this coevolution, we would like to note that, immediately after duplication of ARs, there is a strong constraint to conserve both autoregulatory binding sites on DNA. To explain this, we can reason as follows. Since the length of a gene is larger than those of a TF binding site, initially neutral mutation is likely to affect one of duplicate proteins, say A'. Now, we suppose A' responds to a different stimulus than A, and is therefore a candidate for innovation. Mutations on the binding site for A' along DNA become dangerous because, in absence of mutations of the DNA-binding region on the protein side, the expression of A' still affects A and all its targets (which might be numerous, as in the case of CRP and FNR). Note that, in the very common case of dimeric TFs (9), a simple way to effectively decouple the system is to mutate the protein-binding site that creates the dimer of A'.

4.2 Scoring Method for Transcription Factor Binding Sites

This section provides the technical details on the specificity analysis of CRP and FNR reported in SI Fig. 12. The two lists of binding sites for these TFs were scored with a simple method based on information theory (8). Consider two duplicate TFs, A and A', having N and N' known binding sites along DNA. Each binding site for A is a sequence $\{b_i^j\}_{i=1..n_b}^{j=1..N}$, where $b_i^j \in \{A, T, C, G\}$, and analogously for A', with primed indexes. The histograms for the two lists can be written as

$$p(')_{ib} = \frac{1}{N} \sum_{j=1}^{N(')} \delta(b(')_i^j, b) \ ,$$

where $\delta(k, l)$ is Kronecker's delta, and the primes between parentheses illustrate the same formula for A'. Neglecting finite-size corrections (8), one can write the information content in the two

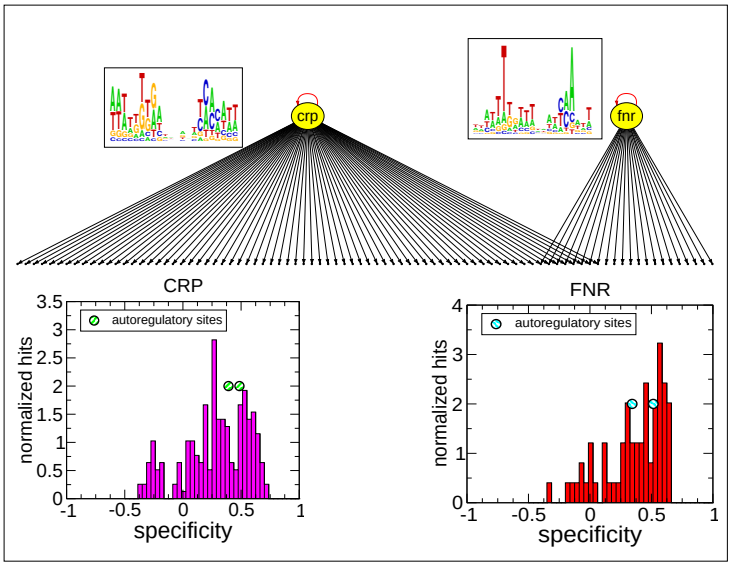

Figure 12: Specificity for CRP and FNR, a likely example of duplicate ARs. Top: logos of the binding sites of the two TFs obtained with the sets of binding sites of the regulonDB data set (computed with no compensation for reverse-complement binding). Middle: scheme of CRP and FNR interactions in the Shenn-Orr data set. The two TFs regulate a large number of targets (73 and 23) and are both found to bind in 8 *cis*-regulatory regions. Bottom: histogram of the specificity of the two TFs, computed keeping into account antisense binding and entropy of mixing of the two sets of sequences. The circles indicate the specificity of the autoregaolatory binding sites of the two TFs.

cases as

$$f_i(') = 2 + \sum_{b\in\{A,T,C,G\}} p(')_{ib} \log_2 p(')_{ib} \quad .$$

The above function measures the low-variability, low-entropy and thus most informative, positions in the list of given sequences. The availability of the two different lists justifies the introduction of a third observable, giving the variability of the total list of available sequences, or the "entropy of mixing" (see, for instance, page 23 in ref. 10). The mixing information content is then

$$f_i^M = 4 + \sum_b (p_{ib} + p'_{ib}) \log_2 (p_{ib} + p'_{ib}).$$

To score for specificity, we use the "specific information"

$$f_i^S = f_i + f'_i - f_i^M.$$

This quantity gives a quantification of the positions in the sequences that are most important to distinguish A from A'.

To score for the specificity of a sequence $\mathbf{b} = (b_1, .., b_{n_b})$, we considered the penalty functions

$$C^{\mathbf{b}}_{A(')} = \sum_{i=1}^{n_b} (i - p(')_{ib_i}) f_i^S \quad .$$

This function describes the "distance" of $\mathbf{b}$ form the histogram of A or A', weighted with the specific information f_i^S relative to each position on the binding site. The specificity of sequence $\mathbf{b}$ for TF A is then defined as

$$S_A(\mathbf{b}) = C^{\mathbf{b}}_A - C^{\mathbf{b}}_{A'}$$

and analogously for A'. To improve the scoring, we considered also the reverse complement sequences (which we indicate by $\bar{\mathbf{b}}$) of the ones available from RegulonDB.

To resume, the above considerations can be used to construct the following algorithm algorithm, that iteratively

- Computes $p(')_{ib}$ and f^S
- Scores for specificity.
- If $S(\bar{\mathbf{b}}) < S(\mathbf{b})$ substitutes $\bar{\mathbf{b}}$ to $\mathbf{b}$ in the list of binding sites.

Note that our considerations do not keep into account explicitly the binding energy of TF-DNA interactions. Biophysical methods (11–13) could possibly yield better results. Moreover, evolutionary models such as the one described in ref. 14, could give more insight in the assessment of the divergence process of duplicate TFs.

5 Extension of the Analysis to the RegulonDB 5.5 Data Set

We report in this section some results obtained using the more recent RegulonDB 5.5 data set (7), containing many more interactions than the Shen-Orr data set used for the main analysis. The newer data set is divided into two sets of interactions, purely transcriptional, and Sigma-factor-like.

5.1 Analysis of Feedback and Hierarchy

For the topological analysis, we considered, the interactions between operons of both the transcriptional set (648 nodes, 147 TFs, 1170 interactions, 85 ARs) and the complete set including also Sigma factors (808 nodes, 153 TFs, 1869 interactions interactions, 88 ARs). For comparison, the Shen-Orr data set has 423 nodes, 117 TFs, 59 ARs, and 578 interactions. Unlike the Shen-Orr data set, both RegulonDB 5.5 data sets contain some non-self-regulatory feedback in addition to abundant self-regulatory links. In particular, they contain, respectively, 4 and 7 pairs of mutually regulating TF pairs. These cross-regulating pairs, all of which contain at least one AR, can be identified in SI Fig. 13. More generally, we have quantified the significance of non-self-regulatory feedback and hierarchy of each RegulonDB 5.5 data set by comparing them with randomized null networks with the same degree sequence, *i.e.* conserving the number of incoming and outgoing links for each node (SI Appendix Sec.1).

As for the Shen-Orr data set, the importance of non-self-regulatory feedback was quantified by the size of the regulatory core obtained after pruning the tree-like input and output cascades using the leaf-removal algorithm LRb (see Fig. 1b and SI Appendix Sec.1). For both RegulonDB 5.5 data sets, this analysis confirms that the amount of transcriptional, non-self-regulatory feedback is significantly smaller than what is typically expected from their null model counterparts, SI Fig. 13.

The importance of hierarchy was also quantified for both RegulonDB 5.5 data sets. As mentioned in the main text, there is no straightforward general definition of hierarchy for networks including feedback. We have used the same definition for the number of hierarchical layers as in the Shen-Orr data set analysis. It is based on the number of iterations of the leaf-removal algorithm, which prunes the tree-like input and output cascades of the regulatory networks. This yields respectively 7 and 6 hierarchical layers for the RegulonDB 5.5 data sets, without and with Sigma factors respectively, against 5 layers for the Shen-Orr data set. Comparisons with null models were restricted to randomized networks with the same regulatory core size as the empirical regulatory network for each RegulonDB 5.5 data sets. As for the Shen-Orr data set, the number of hierarchical layers for either RegulonDB 5.5 data set was found to be remarkably lower than its respective randomized network counterparts, SI Fig. 14. Yet, although significant, these results only compare by definition the hierarchical structures of the tree-like input and output cascades of the empirical networks versus randomized counterparts. Unlike in the Shen-Orr empirical network, both RegulonDB 5.5 empirical networks and all their randomized counterparts exhibit a finite core with non-self-regulatory feedback, which is *de facto* excluded

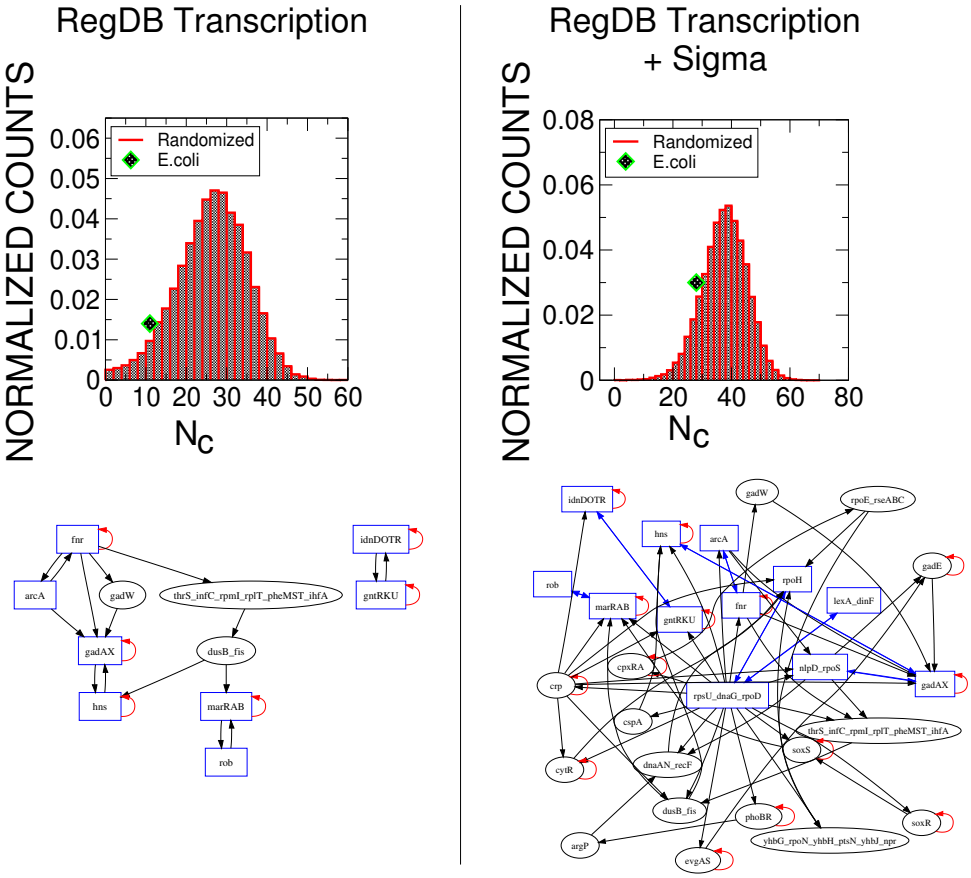

Figure 13: Feedback in the RegulonDB 5.5 graph and its randomized counterparts. The left panel refers to purely transcriptional interactions, while the right panel contains results for the set including also Sigma factor interactions. Top: Comparison of the histogram of the LRb core size N_c (number of nodes in the core) of 10^5 randomized counterparts with the empirical networks. Both panels show that the amount of transcriptional, non-self-regulatory feedback is significantly smaller than typically expected from null model counterparts. Bottom: LRb core for the two networks. Self-regulatory links are in red, and nodes involved in two-node feedback loops are blue rectangles.

from the adopted measure of hierarchy that focuses on the tree-like input and output network components only.

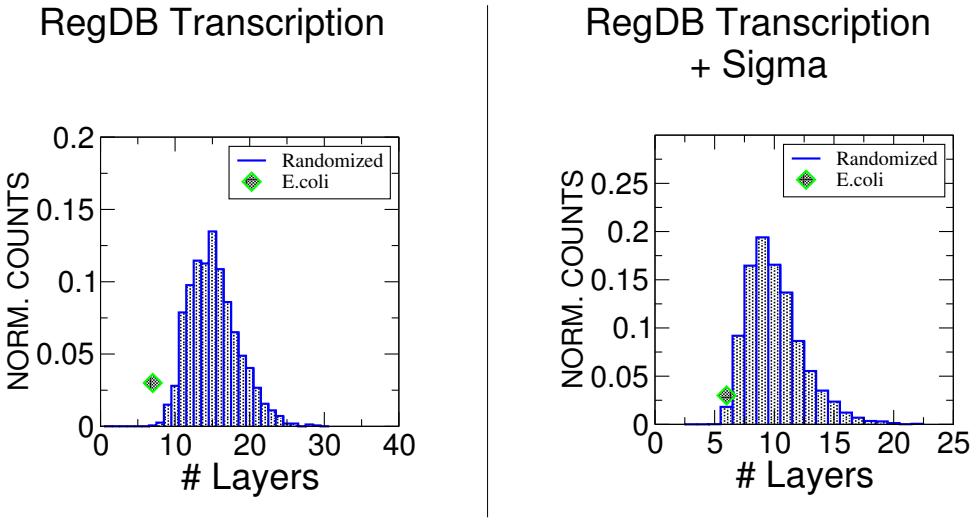

Figure 14: Histogram of the number of layers defined as total leaf-removal iterations for randomizations of the RegulonDB 5.5 transcriptional interaction graph (left) and the graph including also Sigma factor interactions (right), compared to the original networks. The randomizations considered for this evaluation have the same core size as the empirical networks.

An alternative definition of hierarchy, which includes *a priori* all network nodes, is based on the length of the longest elementary (*i.e.* without feedback loops) path between all pairs of root and leaf nodes of the directed graph. This is a natural extension of the definition of hierarchy on tree-like directed networks, which also ignores the role of self-regulatory links. Note that both hierarchy definitions are strictly equivalent on tree-like networks and thus yield 5 hierarchical layers for the Shen-Orr empirical network. By contrast, defining network hierarchy in terms of the longest elementary path for both RegulonDB 5.5 data sets yields 7 hierarchical layers for the transcriptional data set and 10 for the complete data set including Sigma factors (as compared to 7 and 6 layers respectively with the leaf-removal definition of hierarchy). Although the number of hierarchical layers can also only increase for the randomized network counterparts, this is computationally difficult to estimate precisely. In fact, finding the longest elementary path on a general non-tree-like graph is an NP-complete problem (as it can be transformed into the hamiltonian path problem) (15). Thus, it becomes exponentially difficult to solve for the large core size of many randomized network counterparts, as shown in SI Fig. 13. Hence, the statistical significance of these longest elementary path results is difficult to estimate in general. Note, however, that the number of layers for RegulonDB 5.5 data set without Sigma factors has remained unchanged at 7 layers, whereas it has necessarily increased for its randomized counterparts. This implies that the difference in hierarchy between empirical and randomized networks for this data set is actually even larger in terms of longest elementary path than in terms

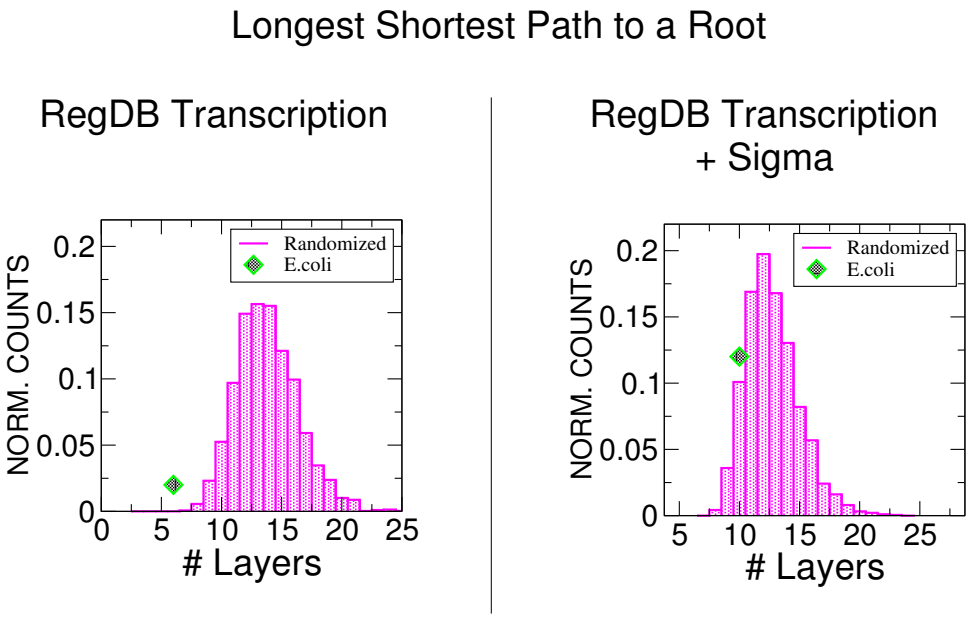

Figure 15: Histogram of the number of layers, defined as the longest shortest elementary path between all root and leaf pairs, for randomizations of the RegulonDB 5.5 interaction data as compared to the empirical networks. The randomizations considered for this evaluation have no constraint on the core size.

of longest input and output branches, assuming the same fixed core size for both empirical and randomized networks.

Overall, this demonstrates that *E. coli* empirical network has indeed a shallow structure with a limited number of hierarchical layers with respect to its randomized counterparts. A somewhat complementary measure of this fact is given by the length of the maximal shortest elementary path between pairs of root and leaf nodes in the empirical versus randomized networks. By contrast with the NP-completeness of the longest elementary path problem, shortest elementary paths can be found in quadratic time using Dijkstra's algorithm (16). This enables to quantify the statistical significance of the "longest shortest elementary path" with respect to randomized network counterparts with large regulatory core size. This quantity is also found to be significantly smaller for the empirical network than for randomized counterparts, particularly in the case of purely transcriptional interactions, SI Fig. 15.

5.2 Evolutionary Analysis

Although the set of interactions provided by RegulonDB 5.5 is a considerable extension of the Shen-Orr data set, the main limitation for the comparison of homology classes and transcriptional interactions at the base of our analysis is the number of genes for which a structural domain sequence is known. With the SUPERFAMILY 1.61 data (6), 616 domain structures are given

for the genes. The 1.69 version of the database, with 844 hits, does not enlarge substantially the sample. With this caveat, the results given for the Shen-Orr data set are all recovered qualitatively, with little quantitative variation. We refer below to results obtained for the transcriptional interaction graph without Sigma factors.

a

	E. coli	Randomized	P
AR class variability g_{AR}	5.571	4.263 ± 1.208	0.14
AR fraction h_{AR}	0.842	0.755 ± 0.041	0.015
Homolog pairs in same layer	1171	420 ± 34.7	1e-4

b

	Genes in network	Transfers	P
TF	150	25	2e-4
TG	1187	372	0.11

Table 1: Evaluation of different evolutionary drives using data from RegulonDB 5.5 (see Table 1 in the main text). (a) Tendency of duplicates of ARs to retain their self-links measured by h_{ar} (average fraction of ARs in classes with two or more ARs), and g_{ar} (variance of the AR population among classes), and preservation of the layer structure, measured by the number of homolog pairs occupying the same layer. (b) Evaluation of the repartition of gene gains by horizontal transfer among TFs and TGs.

Duplication. For ARs, one recovers the signatures of duplication (Table 1a). In this case, most of the signal comes from the global parameter h_{ar}, compared to g_{ar}, meaning that the increased AR population in the classes dominates over the variability among classes. Despite the presence of two pairs of homologous genes with crosstalk (gntR,idnR; rob,marA), the self-links are still the large majority of the links within homology classes (being 89% of the total). The total number of links within the same homology class is 8.

Layer Hierarchy. With the definition given above, i.e. gene i lays in layer l if the longest open path upstream pointing to i involves $l-1$ different nodes, we still find a strong signal for the tendency of genes of the same class to occupy the same layer (Table 1a).

Horizontal Transfer. Finally, also within this data set, comparison with lists of imported genes obtained with different phylogenetic tree reconstruction algorithms shows the tendency of gene gains be TGs and to be expelled from homology classes (Table 1a).

References

[1] Cosentino Lagomarsino, M., Bassetti, B. & Jona, P. (2006) *Randomization and Feedback Properties of Directed Graphs Inspired by Gene Networks*, Lecture Notes in Computer Science, ed. Priami. C., . (Springer Berlin / Heidelberg) Vol. 4210, pp. 227–241.

[2] Cosentino Lagomarsino, M., Bassetti, B. & Jona, P. (2005) in *Soft Condensed Matter: New Research*. (Nova Science Publishers). (q-bio.MN/0502017).

[3] Cosentino Lagomarsino, M., Jona, P. & Bassetti, B. (2005) *Phys Rev Lett* **95**, 158701.

[4] Gough, J., Karplus, K., Hughey, R. & Chothia, C. (2001) *J Mol Biol* **313**, 903–19.

[5] Madan Babu, M. & Teichmann, S.A. (2003) *Nucleic Acids Res* **31**, 1234–44.

[6] Teichmann, S.A. & Babu, M.M. (2004) *Nat Genet* **36**, 492–6.

[7] Salgado, H., Santos-Zavaleta, A., Gama-Castro, S., Peralta-Gil, M., Penaloza-Spinola, M.I., Martinez-Antonio, A., Karp, P.D. & Collado-Vides, J. (2006) *BMC Bioinformatics* **7**, 5.

[8] Schneider, T.D. (2002) *Appl Bioinformatics* **1**, 111–9.

[9] Ptashne, M. (1992) *A Genetic Switch*. (Cell Press, MA), Second edition edition.

[10] Ma, S.K. (1985) *Statistical Mechanics*. (World Scientific).

[11] Stormo, G.D. & Fields, D.S. (1998) *Trends Biochem Sci* **23**, 109–13.

[12] Djordjevic, M., Sengupta, A.M. & Shraiman, B.I. (2003) *Genome Res* **13**, 2381–90.

[13] vonHippel, P.H. & Berg, O.G. (1986) *Proc Natl Acad Sci U S A* **83**, 1608–12.

[14] Mustonen, V. & Lassig, M. (2005) *Proc Natl Acad Sci U S A* **102**, 15936–41.

[15] Garey, M.R. & Johnson, D.S. (1979) *Computers and Instractability: a guide to the theory of NP-completeness*. (Freeman, New York).

[16] Dijkstra, E. W. (1959) *Numerische Mathematik* **1**, 269–271.