# A comparative evolutionary study of transcription networks. The global role of feedback and hierachical structures†

A. L. Sellerio,[ab] B. Bassetti,[a] H. Isambert*[c] and M. Cosentino Lagomarsino*[a]

*Received 2nd September 2008, Accepted 31st October 2008*
*First published as an Advance Article on the web 25th November 2008*
**DOI: 10.1039/b815339f**

We present a comparative analysis of large-scale topological and evolutionary properties of transcription networks in three species: the two distant bacteria *E. coli* and *B. subtilis*, and the yeast *S. cerevisiae*. The study focuses on the global aspects of feedback and hierarchy in transcriptional regulatory pathways. While confirming that gene duplication has a significant impact on the shaping of all the analyzed transcription networks, our results point to distinct trends between the bacteria, which display a hierarchical network structure with short transcription cascades, and yeast, which seems able to sustain a higher wiring complexity, including larger feedback, longer transcription cascades, and the combinatorial use of heterodimers made of duplicate transcription factors, absent in *E. coli*.

## Introduction

Cells need constant sensing of environmental changes and internal fluxes, and the correct response to these external and internal stimuli through the simultaneous expression of a large set of genes. The basal mechanism that performs this task is transcriptional regulation. Depending on species, context and specific function, this can involve simple interactions or complex signaling cascades, but in general it involves a large number of genes.[1–6] For this reason, it is necessary to characterize this regulatory process from a global, or "network" point of view. To this aim, transcriptional regulation networks are defined starting from the basic functional elements of transcription.[7] This information is represented as a directed graph, usually identifying each gene transcript and their protein products with an unique node, and each regulatory interaction with a directed edge $A \rightarrow B$ between the node $B$ (the target gene) and the node $A$ (the gene coding for a transcription factor (TF) that has at least one binding site in the cis-regulatory region of $B$). A transcription factor regulating its own expression is called an autoregulator (AR). With this definition, the interaction graph structure is given by large-scale and collections of small-scale experiments.[8–11]

A basic way to understand the architecture of transcription networks it to consider their topology. Topological analysis is able to capture functional properties and important architectural features of the network,[12–19] by comparing with suitable null models. For example, "network motifs" are subgraphs for which the probability $P$ of appearing in a random network an equal or greater number of times than in the empirical network is lower than a given cutoff value. For a meaningful comparison, null random networks are taken with the same single-node characteristics of the empirical network. Usually, each node is constrained to have the same number of incoming and outgoing edges (degrees) as the corresponding node has in the empirical network.[20,21] Considering more large-scale properties, the known transcription networks possess a hierarchical feedforward layered structure,[15,17] and often feedback is mainly limited to a rather large set of autoregulations.[22]

The perspective on the topology transcription networks is enriched by taking an evolutionary point of view.[23] Evolution of a transcription network is driven by three main biological mechanisms: (i) gene duplication, (ii) rewiring of edges by mutation/selection of TF/DNA interactions and (iii) horizontal gene transfer. The first mechanism has been shown to play a substantial role, although the extent to which it can shape the network is debated.[7,14,24–26] For example, it has been shown that network motifs do not emerge from duplication events,[7,26] while other topological properties have arisen from gene duplications.[7,27] We have previously considered from this viewpoint the properties of hierarchy and feedback in the *Escherichia coli* network,[19] finding that gene duplication can be held responsible for the preservation of self-regulations, and for the "shallow" layered organization, which one can hypothesize to optimize the time constraints for the production of targets.

In this paper, we present a comparative study of transcription networks, which extends the analysis on *E. coli*, and considers also the evolutionarily distant bacteria *Bacillus subtilis*, and the eukaryote *Saccharomyces cerevisiae*. From comparison of these data we distinguish between unifying and distinct features of the three networks. In particular, while in the two bacteria feedback loops involve few nodes and the number of hierarchical layers is minimal, yeast shows more feedback and a more complex hierarchy of transcription factors. If we take into account the effects of evolution on topology, a richer, more complex scenario appears. Our analysis is focused on the mechanism of gene duplication,

[a] *Università degli Studi di Milano, Dip. Fisica. Via Celoria 16, 20133 Milano, Italy. E-mail: Marco.Cosentino-Lagomarsino@unimi.it*
[b] *Current address: École Polytechnique Fédérale de Lausanne, Institut de Physique de la Matière Complexe, CH-1015 Lausanne, Switzerland*
[c] *UMR 168/Institut Curie, 26 rue d'Ulm, 75005 Paris, France. E-mail: herve.isambert@curie.fr*

† Electronic supplementary information (ESI) available: Supplementary figures and tables. See DOI: 10.1039/b815339f

 

and is based on a network growth model which considers this drive to be the only one present. With this method it is also possible to infer on other mechanisms indirectly. Our main findings are the following: (i) duplications play an important role in evolution of the TN—the relative abundance of simple network subgraphs stemming from duplication is evident in all data sets; (ii) gene duplications shape the degree sequences and the hierarchy of the network, as predicted by a duplication–divergence model; (iii) the feedback core in the yeast network may be shaped by duplications of existing feedbacks; (iv) the yeast network tends to form heterodimeric TF pairs from duplicates, which seem to be forbidden in *E. coli*, possibly because of the same evolutionary constraint promoting crosstalks elimination between AR duplicates.

## Results

### Feedback and hierarchy. Topological evaluation

Feedback is present in a network if closed directed paths exist. We quantify the amount of feedback with the "leaf-removal" algorithm. This decimation algorithm iteratively removes the input and output tree-like components (*i.e.* parts without loops) of a directed graph.[28] The outcome can be either the empty graph or a non-empty subgraph of the original. In the former case, the whole network is tree-like. In the latter, we say that the leftover subgraph represents the feedback *core* of the network. The size of this core (number of nodes or edges, with respect to the complete network) can be used as a measure of the amount of feedback. The core is the set of all and only the nodes involved in feedback loops. Note that this need not be a single connected component.

The reverse, or complement, of feedback is *hierarchy*. Neglecting self-regulations, we define *roots*, the nodes that are not regulated by other nodes, and *leaves*, the nodes that do not regulate other nodes. Starting from the roots of the graph we can define hierarchical layers considering the relative ordering of nodes in a chain of regulatory interactions. (i) If we consider a tree-like graph (or a tree-like subset of it), each iteration of the leaf-removal algorithm removes a number of nodes, either roots or leaves, and the edges pointing from or to these nodes. By definition, the removed nodes do not interact with each other, and can be thought to form a distinct computational "layer". In this case, the leaf-removal algorithm described above naturally defines a hierarchy between the nodes, through the longest path to a root or to the feedback core upstream of a given node. In the general case, we define the number of layers in the hierarchy by the number of iterations of the leaf-removal algorithm needed to remove all the nodes outside of the feedback core (thus as a sum of layers downstream and upstream of the feedback). (ii) Hierarchical layers can be defined in a similar way through the longest open chain of regulators that each of their members share. In a longest-path hierarchy, and in the absence of feedback, members of layer one are regulated by, at most, themselves. Members of layer two are regulated by a chain of one, and no more, nodes and possibly themselves, and so on. This definition, though conceptually similar to that given by the leaf removal, is computationally demanding since the algorithm represents a NP-complete problem. We did not use this definition in our analysis. (iii) An alternative definition (computationally easier) is given by the shortest open directed paths between nodes. The number of layers is computed considering the longest among the shortest paths from any pairs of nodes (which can be found in polynomial time, for example with the Dijkstra algorithm[29]). Shortest paths measure the minimal number of intermediary transcriptional interactions required for a signal from a transcription factor to reach a given target downstream, which can be interpreted as hierarchical layers. On the other hand, a straightforward univoque definition of hierarchical layers using shortest paths is difficult to produce.[17]

In order to quantify the significance of feedback and hierarchical properties of the three networks, we compared the five data-sets with random ensembles of networks having the same degree sequences, *i.e.* conserving the number of incoming and outgoing edges for each node. In particular, we chose not to conserve the number of self-regulators. Our previous work[19,21] shows that in general this null model yields qualitatively different results. In particular, in the case of the Shen–Orr data set the number of layers falls in the average of the random ensemble that conserves self-regulations. Here, we hypothesized that, as in the model for the graph growth by duplication, auto-regulatory edges have the same status of other edges, and we generally did not consider this kind of randomization. For all the bacterial data sets, we consistently find the following: the feedback core is slightly smaller than the typical random case, and the number of layers is minimal. This is true both for a longest path and a shortest path hierarchy. Data are shown in the first two columns of Fig. 1. On the other hand, the case of *S. cerevisiae* is more complex. First, we observe that the two data sets we analyzed are not consistent with each other. We believe this to be due to their large differences in size: the older Guelzim data set[30] contains approximately 800 interactions, and is a subset of the more recent Balaji data set,[27] which contains around 13 000 interactions. We considered the Guelzim set as a separate case because it is still used in recent analyses.[26] However, as we will see, for some of the analyses we performed this data set revealed to be too small. In this case, its nearly tree-like topology does not differ from randomizations in any of the observables considered above, which is a different trend from the bacteria, but also points to possibly insufficient statistics. For the larger Balaji data set, we find marked contrasting trends compared to the bacterial networks. The rather large feedback core involving 60 nodes and 207 interactions is far larger than the one found in the two bacterial networks, and made of a large connected component. On the other hand, it is significantly smaller than the typical feedback core found in randomizations. Interestingly, the number of leaf-removal layers falls in the average, in contrast with the trend observed in bacterial networks. Even more surprisingly, the number of shortest-path layers in the yeast network greatly exceeds randomizations, in strong contrast with the bacterial data-sets. Data are shown in the third column of Fig. 1. In more detail (Fig. S1†), the lengths of these shortest-paths have a Poisson-like distribution in both the empirical graphs and their randomizations, but for the yeast network the tails of

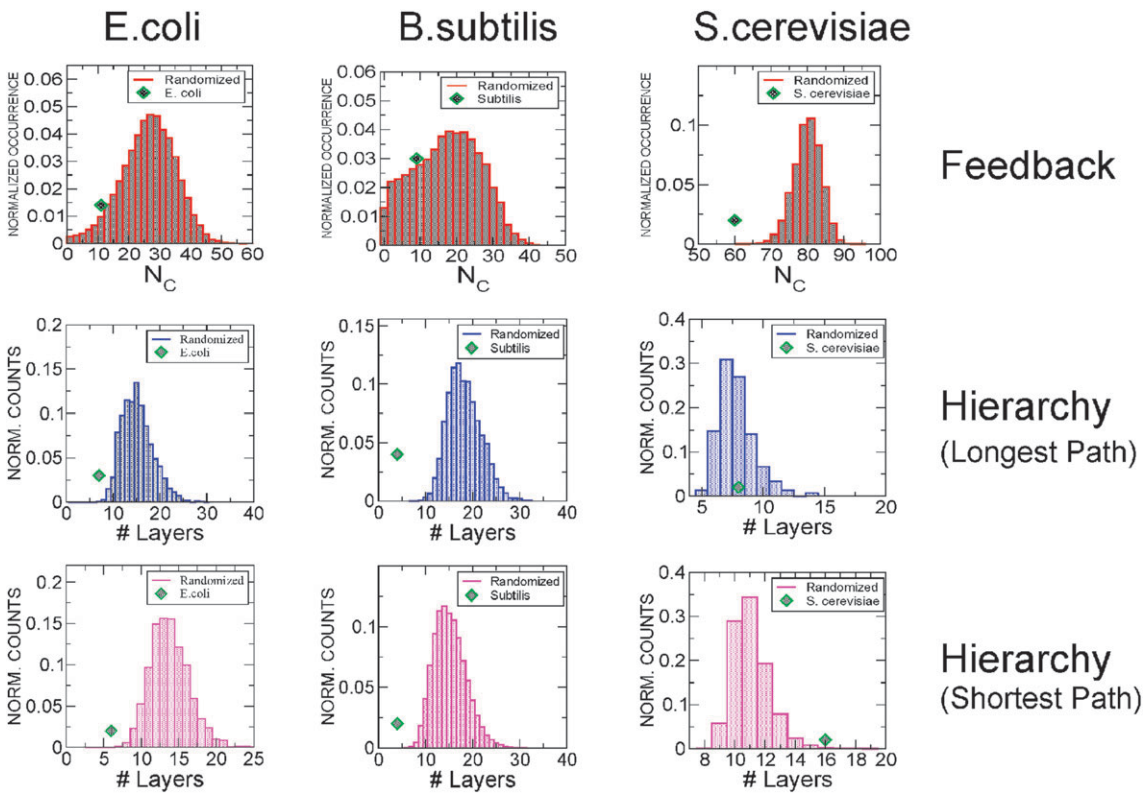


**Fig. 1** Feedback and hierarchical layers for the three networks of *E. coli* (RegulonDB 5.5) *B. Subtilis* (DBTBS) and *S. Cerevisiae* (Balaji data set), compared to randomized instances. The data on *E. coli* are compatible with older data sets. The two data sets on yeast do not give compatible results; data refer to the most recent one. Top: number of nodes, $N_C$, in the feedback core. Middle: number of leaf-removal layers for the empirical networks and randomizations having the same or lower $N_C$ as the empirical ones (to improve the size of the sample, we show $N_C < 70$ for the case of yeast, however, the results do not change for lower thresholds). Releasing the constraint on $N_C$ gives a weaker signal for the bacteria. Bottom: number of layers in a shortest-path hierarchy, without constraints on $N_C$ for the randomized cases. A constraint for the empirical $N_C$ gives a stronger signal.

their distributions are significantly shifted between the two. Note that the results hold also in presence of artificially generated noise in the data, by substituting up to 10% of the interactions with uniformly chosen random ones (see Fig. S2†). They are also stable for randomizations that conserve self-regulations.

### Evolutionary duplication–divergence growth model

We use a simple duplication–divergence model for the growth of the transcription network[19,31,32] to guide the data analysis. The model is formulated as follows. At each step all, or a fraction of the nodes are duplicated. The duplicate nodes inherit all the in- and outgoing edges of the original (ancestral) node. In other words, one supposes that before divergence the binding sites on the proteins and the regulatory regions on DNA remain identical. Subsequently, the new edges are removed with a certain probability, that might depend on the status of the node (regulator or target, subject or regulating a new or an old gene, *etc.*)

A qualitative study of the model leads to the following schematic results.[31] First, by definition there is no possibility of *de novo* addition of edges by rewiring. Hence no edges can be created from an homology class that initially does not regulate another one. In the simplest case, if feedback is initially absent, it cannot arise spontaneously. Moreover, no hierarchical layer can be added to the network, and duplicates selected for fixation will lie in the same layer. On the contrary, in the presence of feedback, and ARs in particular, the model behaves differently. Duplication of ARs can give rise to higher order feedback and to new ARs. However, this feedback will be strictly confined among members of the same homology class. Analogously, the degree distribution is piloted by the degree distribution of duplicates. The in- and out- degree distributions can be decoupled by choosing different removal probabilities for old/new and new/old edges. One can obtain power-law out-degree and compact in-degree distributions that are in close relation with the phylogenetic conservation of the network nodes:[32,33] conserved nodes *must* exhibit an in- or out-degree sequence with a scale-free distribution, while nodes with exponentially distributed degrees *cannot* be phylogenetically conserved under a general duplication-divergence model. Thus, according to this model, bacterial transcription networks with scale-free out-degrees and exponential in-degrees are consistent with a phylogenetically conserved set of transcription factors and a non-phylogenetically conserved set of target genes (possibly due to abundant horizontal tranfers of target genes[34]).

In a realistic situation, the above observations may not apply because of the possibility of edge rewiring, that may be able to shuffle the hierarchy, create new feedback, affect the degree

distribution, and mix edges among homology classes. Note, however, that some regulatory edges between apparently different homology classes may have also resulted from very old edges within ancestral homology classes that are no longer classified as single homology classes due to their gradual divergence. Besides, while actual edge rewiring *may* have occurred after homology class separation, it is clear that edge deletion *must* typically happen after gene duplication (or else extant regulatory networks would be densely cross-edged graphs.) Hence, this model, which focuses on the necessary deletion of duplicated edges, can assess the specific role of gene duplication and edge deletion in the growth of the network and also underline their possible shortcomings to delineate the role of other evolutionary processes such as edge rewiring.

### Data evaluation

Having in mind the above qualitative results, we examine the topological roles in empirical networks of nodes coming from the same common ancestor. We define proteins that are likely to share a common ancestor through structural domain composition.[14,35,36] We define homologs as proteins whose domain architectures (ordered sequence of domains) are identical neglecting domain repeats. This corresponds to a conservative view of homology where no new domains are acquired or lost after duplication. The results are tested using different definitions of homologs.[37] Note that in this analysis and in subsequent ones we assume homology is mostly detecting duplications, while paralogs could also arise from horizontal transfers. This assumption will be (statistically) valid if horizontal transfer is a relatively minor drive in the shaping of the genome (as it is the case in yeast), or in any case if the horizontally transferred nodes occupy a peripheral role in the network, as we verified in *E. coli* (see the Discussion for further comments on this point).

We have analyzed the distribution of regulatory edges between and within classes of the likely duplicate genes. The statistical significance of the analysis in terms of homology classes is established[7] by comparison with random shuffling of genes (TFs and TGs separately) between classes. This analysis is limited by the number of nodes for which homology classes can be constructed, and thus less sensitive to the particular data set. In all cases we find that the motifs of duplicated TGs, regulated by a single or duplicate TFs, and duplicate TFs regulating a common target (Fig. 2), are significantly over-represented,[14] which can be seen as a validation of the model. The signal for this is smaller in the smaller data sets (such as *B. subtilis*), because of poor statistics.

### Degree sequences and duplications

A first question to address is whether gene duplications are able to pilot the observed degree sequences of the network, as predicted by the model. For the case of yeast, one can perform this test on the above-defined homology classes, and also on the doubly conserved genes in the yeast whole-genome duplication found by Kellis *et al.*[38] We measured the degree distributions restricted to targets or regulators falling in the same homology class, and verified whether they scaled in a similar way as the unrestricted distributions.

Our results, considering the Balaji data set for yeast and RegulonDB 5.5 for the *E. coli*, are shown in Fig. 2. The homologs follow a qualitatively similar distribution to the entire network, exponential for the in-degree, and broader (compatible with a power law with cutoff) for the out-degree. This confirms the general idea that topological features as the degree distributions are strongly dependent upon duplications. The same observation holds for measures of hierarchy (Table S2†).

In the case of yeast, the out-degree of the entire network seems to follow a broader distribution than that of the duplicates (top-left panel of Fig. 2). This can be seen as an indication that other processes, such as rewiring, concur in defining the targets of a TF. The behavior of the duplicates from the whole-genome duplication is conditioned by the small size of the sample, and its degree distribution drops with a faster decay for both the in- and the out-degree.

### Duplication of autoregulatory circuits

The role of autoregulatory interactions appears to be rather different in the bacterial data sets compared to yeast. The fraction of self-regulated TFs is above 50% in all three bacterial data sets, while being around 10% only in the two *S. cerevisiae* data sets. If we measure directly the distribution of ARs in classes, in the *E. coli* datasets we find a consistent signal that the population of ARs in homology classes is more dense and more variable ($Z$ score $\simeq$ 2–3) than in randomized instances.[19] This is a direct evidence of duplication and conservation of ARs, and agrees with the conservation of a hierarchical structure during evolutionary growth (see also Table S2†). In *B. subtilis* and yeast, we find that the population of ARs in classes falls closer to (although systematically above) the average of the randomized samples ($Z$ between 0.6 and 1).

On the other hand, in all the data sets we find evidence for the relevance of AR duplications if we consider the contributions of all the subgraphs that can stem from AR duplication (Fig. 3) within homology classes. In order to quantitatively evaluate the relevance of the subgraphs, we define an observable that counts the occurrence of the following events

1. An AR regulates a homologous, non-self-regulatory transcription factor.

2. An AR is regulated by a homologous, non-self-regulatory transcription factor.

3. Two homolog ARs possess reciprocal cross regulation.

We indicate with $N_C$ the total number of homology classes, with $M_i$ the total number of the subgraphs of the three kinds found in homology class $i$, and by $L_i$ the number of edges in the class, and define

$$a = \frac{1}{N_C}\sum_{i=1}^{N_C}\frac{M_i}{L_i}.$$

This observable represents the ratio of subgraphs to the total number of (self- and non-self-) edges present in the homology class. We find (Table 1) that the AR duplication subgraphs are systematically over represented. Specifically (see Table S1†), in *E. coli* and *B. subtilis*, this signal is dominated by duplicated self-interactions followed by loss of crosstalks, while for the

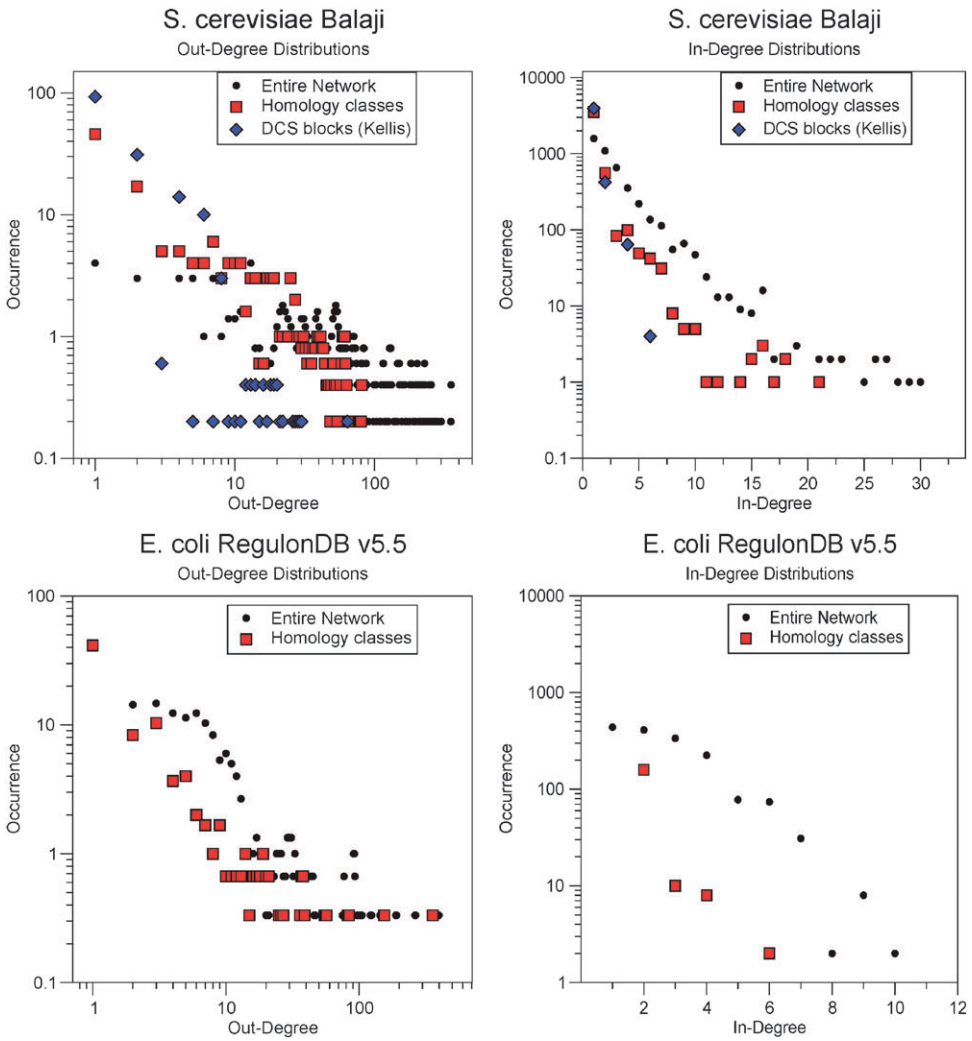


**Fig. 2** In- and out-degree distributions of *S. cerevisiae* (Balaji data set) and *E. coli* (RDB 5.5), compared to those obtained from homolog TGs and TFs. Homology is assessed with domain-architecture homology classes, and in the yeast, through doubly conserved genes in whole genome duplication. The in-degree shows good agreement (right panel) with the prediction of the duplication model. For the out-degree distribution, instead, data for the *S. cerevisiae* (top left) for the whole network seems to be wider than the homology classes. Some histograms have been smoothed out by averaging nearby bins in order to enhance the visibility of the tails.

*S. cerevisiae* network, subgraphs including crosstalks are more abundant, compatibly with the observed more complex hierarchy. In addition, in the significantly overrepresented AR duplication subgraphs (type *ii* in Fig. 3), only one of the two possible self-regulations of duplicate TFs survives, indicating that ARs do not proliferate by duplication. This observation indicates that duplications have been important to shape the non-self-regulatory edges within homology classes, thereby redefining the TF hierarchy, besides leading to the fixation of new ARs in the bacterial networks. This is especially the case in yeast, indicating that circuits with crosstalks or feedback stemming from AR duplication are not negligible, and might have contributed to the build-up of the existing feedback.

### Interaction across homology classes and evolution of the feedback core

To gain further insight into this point, we evaluated the distribution of interactions within and across homology classes. Considering the transcription factors homology classes, we constructed a "collapsed" weighted network as follows. The nodes are homology classes, and weighted oriented edges represent the number of members of a class regulating members of another class. In absence of rewiring, this would be a likely "ancestral" TF–TF network, inherited by duplication of "primitive" self- and heterologous edges. The duplication model dictates that this ancestral graph can only be hierarchical or maintain the same amount of feedback of the initial configuration. In particular, if an ancestral edge is sent out from a class to a second one and the reverse is not true, there will be no edges coming from the second class to the first in the evolved network. This corresponds to an *asymmetric* regulatory control of the first homology class over the second homology class. Conversely, the appearance of symmetric regulations may be a signature of rewiring, if one assumes that retaining both ancestral crosstalks produced by the duplication of an ancestral AR was then as unlikely as it appears to be between recent AR duplicates. Hence, symmetric regulation between homology classes in the extant

 

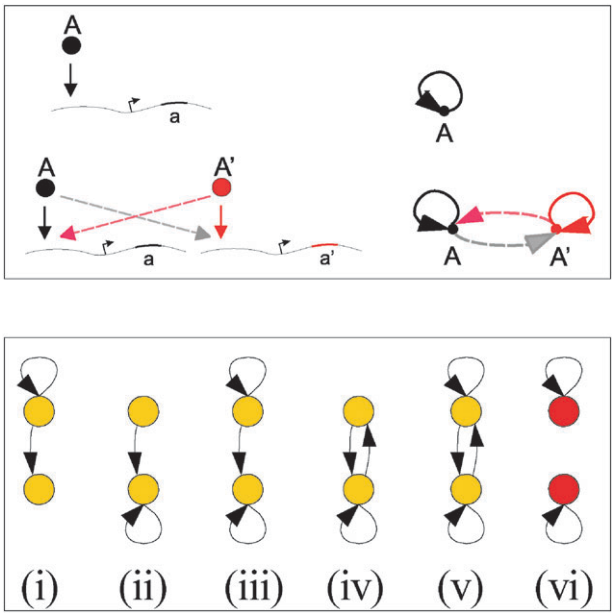


**Fig. 3** Top: scheme of the possible circuits stemming from the duplication of a transcriptional self-regulation after partial inheritance of interactions. Bottom: network subgraphs compatible with AR duplication and inheritance of crosstalks within homology classes, quantified in Table 1 and Table S1.† The circles represent homologous nodes.

**Table 1** Subgraphs compatible with AR duplications, containing crosstalks (*i*–*v* in Fig. 3). The observable *a* measures the ratio between subgraphs and edges in the network, and is defined in the text

| *a* | Empirical | Randomization | *Z*-score |
|---|---|---|---|
| *E. coli* (RDB 5.5) | 0.23 | 0.10 ± 0.082 | +1.6 |
| *B. subtilis* | 0.03 | 0.0042 ± 0.0073 | +3.55 |
| *S. cerevisiae* (Balaji) | 0.13 | 0.020 ± 0.039 | +2.79 |

network actually suggests that the likely initial asymmetry conserved by duplication was progressively smoothed down by edge rewiring or other evolutionary mechanisms with an equivalent effect.

We define the observable $R$, which represents a measure of the strength of asymmetric regulation across the classes. We indicate as $L_{ij}$ the number of regulatory edges between homology class $i$ and class $j$. Let then $C_i$ represent the number of elements in homology class $i$. Then we define

$$R = \sum_{i \neq j} \frac{|L_{ij} - L_{ji}|}{\sqrt{C_i^2 + C_j^2}}$$

where the sum is performed over all the pairs of homology classes $i$ and $j$, without considering self-interactions of classes.‡

‡ The normalization factor in this formula is chosen to be a linear function of the number of edges. The linear relation is suggested by the experimental observation that the number of edges in these TF–TF networks is comparable to the number of nodes. We introduced the square root normalization factor in order to compensate for the bias generated by the different size of classes $i$ and $j$. The important fact here is that the normalization factor should compare to the number of nodes, as does the total number of edges in the network, alternative choices such as $C_i + C_j$ or $sqrt(C_iC_j)$ lead to the same results.

**Table 2** Rewiring parameter $R$ for the TF–TF networks. For the bacterial datasets this observable falls in the typical case of a randomized instance. In the case of *S. cerevisiae* there is a significant trend which indicates absence of rewiring (positive Z score), despite of the larger degree of feedback. This indicates that the feedback circuits tend to lie within homology classes. The parameter $R$ quantifies the strength of asymmetric regulations between homology classes and is defined in the text

| *R* | Empirical | Randomization | *Z*-score |
|---|---|---|---|
| *E. coli* (RDB 5.5) | 9.87 | 7.71 ± 3.65 | +0.59 |
| *B. subtilis* | 3.32 | 3.46 ± 2.64 | ≈0 |
| *S. cerevisiae* (Balaji) | 15.63 | 10.72 ± 2.05 | +2.39 |

Note that the collapsed network condenses in a node all feedback likely inherited by AR duplication. To understand this, it is sufficient to see that all feedback generated by duplication–inheritance of an original self-regulation has to involve the same homology class (and thus the same node of the collapsed network), while feedback generated by rewiring has *a priori* the possibility to link any TF-target pair. In other words, $R$ measures the extent to which feedback is retained within homology classes (and thus likely comes from AR duplication, or duplication of existing higher-order feedbacks between homologs). To sum up, the larger is $R$ (compared to randomizations), the less likely edge rewiring or equivalent mechanisms have shaped the feedback in the network, and the more the hierarchy between the homology classes is maintained.

Evaluation of empirical data (Table 2) shows that in the bacterial data sets, the estimated rewiring is consistent with, or slightly lower than, typical randomized values. Perhaps unexpectedly, we find an even stronger signal for the yeast TF–TF network, where, as we have shown, the feedback is more common. Most of this feedback involves homolog TFs, and the network is likely to have been shaped by a small amount of rewiring, or by rearrangements that keep into account the homology relations of transcription factors. Note that, empirically, one alternative possibility to find this result is that a duplication of ARs with inherited crosstalk is followed by a subsequent split the homology class, so that the two homologs cease to be classified as such. This phenomenon would lead to a false positive for rewiring in our interpretation. However, this problem should not be relevant in presence of a negative signal, such as the one we find.

### Duplication of homodimeric and heterodimeric TFs

An issue where we found notable evolutionary differences between the evolutionary transcriptional architecture of *E. coli* and *S. cerevisiae* is the duplication of homo- and heterodimeric TFs.§

It is known from sparse observations that yeast tends to use more heterodimeric TF pairs than prokaryotes.[39–41] A systematic analysis using large protein interaction datasets confirms this trend. In *E. coli* (out of 150 TFs) we find 21 homodimers, and only 5 heterodimer pairs, all formed with the histone-like protein HU, which has a rather special status. Though no systematic data is available, the common opinion

§ This analysis was not possible for *B. subtilis* due to the lack of large-scale interaction data.

**Table 3** A. Homodimer heterodimer TFs co-occurrence in classes measured by the product of the number of homodimers and heterodimers in each class, summed over classes. B. Homologous heterodimer TFs in the same class

| | Co-occurrence | Empirical | Randomization | *Z*-score |
|---|---|---|---|---|
| A | *E. coli* | 0 | $0.46 \pm 1.22$ | +0.37 |
| | *S. cerevisiae* | 84 | $32.38 \pm 18.95$ | +2.72 |
| B | *E. coli* | 0 | $0.213 \pm 0.45$ | +0.47 |
| | *S. cerevisiae* | 15 | $4.86 \pm 2.29$ | +4.43 |

is that these figures likely underestimate the number of homodimers. A computational evaluation of the TFs available in the 3DComplex database[40] is not able to enrich the sample. On the other hand, it shows that 37/42 TF entries in the PDB are scored as homodimers, corroborating the common belief. Conversely, in yeast, where more systematic data is available, we find (out of 157 TFs) 45 homodimers, and 91 heterodimer pairs.

Inspection of the homology classes (Table 3 and Supplementary Fig. S3†) shows that in *S. cerevisiae*, heterodimeric TFs tend to cluster in homology classes, and also to co-occur with homodimers. We interpret this as a strong indication that heterodimers stem from duplications of other dimeric TFs. This is measured by evaluating the overrepresentation of number of homodimers and heterodimers in homology classes, and the product of homodimers and heterodimers in the same class. The same process seems to be forbidden in *E. coli*. This is possibly related to the evolutionary trend of crosstalk elimination between AR duplicates. Due to insufficient data, we could not show that homodimeric TFs in this bacterium are likely to form classes of duplicate homodimers.

## Discussion and conclusions

One of the main limitations in comparing transcription networks at large scales is the fact that the statistically relevant datasets are still very few in number, so that necessarily one has to compare evolutionary distant organisms, and cannot follow the standard methods of comparative genomics. Inevitably, with the availability of new and more reliable databases in the future, these issues might have to be re-examined. Nevertheless, the parallel analysis is instructive to evaluate major differences or common trends associated to large evolutionary transitions, and to establish methods for future analyses. While previous comparative studies have established the role of gene duplication in shaping regulatory network,[14] the role of self-regulations,[22] and the convergent evolution of existing small regulatory circuits,[14,24,26] this is the first comparative network study focused on the global properties of feedback and hierarchical structures.

Our results highlight both common and distinct trends in the two bacteria and the yeast, both in the topology and in its evolution. We find that the underrepresentation of feedback is common in the three transcription networks. Indeed, the feedback core is always smaller than what expected from the null model, in both the two bacteria and yeast. On the other hand, there is a difference in the size of the feedback core of the network, which is 6 times larger in yeast, involving a major fraction of the transcription factors. As noted by Jeong and Berman,[42] this feedback essentially condenses in one single connected component, and is enriched with TF nodes having the endogenous functions of cell cycle and sporulation, and the exogenous functions of diauxic shift and DNA repair. However, it must be noted that this feedback component, though large, is much smaller than expected by a degree-sequence-conserving null network model, so that the trend of underrepresentation of feedback has to be regarded as a common feature of the networks, and is particularly strong in yeast. On the other hand, there is a remarkable difference in the hierarchical organization between the two evolutionarily distant bacteria and yeast, which emerges more prominently in the length of shortest paths in the network.

In the bacterial networks, the nodes are organized in a small number of hierarchical layers pointing to a minimization of both longest and shortest paths between TFs and their targets. By contrast in yeast, shortest paths are much longer, both in absolute terms and compared to random networks. This presumably reflects different evolutionary constraints for bacteria and yeast and possibly other eukaryotes. One hypothesis could involve differences in transcription time constraints,[43] which are expected to favor short transcriptional cascades against long ones. It is known that generally bacteria grow faster than yeasts, and also that self-repressed transcription factors, which reach steady-state expression in shorter times.[43] However, other specific characteristics of prokaryote and eukaryote cells may also play a role. For instance, we can speculate a role of the differences in post-transcriptional regulation control and in the specificity of the degradation machinery.

We also would like to observe that, while the distributions of small network motifs seem to be common among known transcription networks,[12] the nonlocal observables considered here, which evaluate the feedback and hierarchichal organization of transcription programs, point to consistent differences in network architecture that accompany the transition from prokaryotes to eukaryotes. Consistently with our results, a recent experimental study using added rewired interactions to the *E. coli* network[44] concludes that the behaviour of small modules is affected to a large extent by the rest of the network in which they are embedded, while position in the network hierarchy correlates with expression.

From the evolutionary analysis, we find in all cases significant indications that the existing feedback stems from lower-order feedback by gene duplication and inheritance of interactions. In particular, AR duplication is significant for all data sets. However, circuits with crosstalks or feedback stemming from AR duplication are found in yeast and to a certain extent in *B. subtilis*, but less in *E. coli*. On the contrary, we observe that crosstalk conservation in *E. coli* is not frequent. Direct measurement of ARs in *E. coli* show a consistent signal that the population of ARs in homology classes is more dense and more variable than in randomized instances.[19] This is a direct evidence of duplication and conservation of ARs during evolution. In *B. subtilis* and yeast, the population of ARs falls closer to the null model. The pure analysis of AR population in homology classes is not sufficient to assess that ARs evolve from duplication and inheritance of self-edges.

The other face of the medal is that outside of the small feedback cores, a hierarchical organization where computational layers are built by gene duplication is visible. More in detail, in *E. coli* and *S. cerevisiae* duplicates tend to populate the same layer, indicating conservation of hierarchy. The same seems to be true for *B. subtilis*, but the signal is weaker. A possible interpretation is again that, in bacteria, minimization constraints on the time-scales for the production of target genes, if present, would translate into selective pressure for the reduction of the number of computational layers. On the other hand, such pressure could be released from yeast due to more efficient post-transcriptional control of gene expression. Our current work explores the hypothesis that part of this higher complexity stems from the known whole-genome duplication event.[45,46] The experimental study on artificial *E. coli* rewiring mentioned above[44] shows that most added network connections do not cause large phenotypic changes, indicating that there is a great evolutionary potential for the cell by the plasticity of its regulatory interactions. It would be interesting to explore the analogous large-scale experiments on artificially induced duplications,[47] in order to compare with these findings.

We should note that the model behind our analysis is based on duplication divergence only, while, as we discussed, at least two parallel processes are found to be relevant for shaping transcription networks, rewiring of interactions[48–50] and horizontal gene transfer.[51] These processes have different relevance for yeast and bacteria. In bacteria, horizontal transfer is found to be very relevant, while rewiring is generally considered relatively less important, although it must play a role in establishing convergently evolving network structures, such as network motifs,[24] and upstream motif flexibility has been reported.[23,52] In a previous study,[19] we have found that horizontally transferred nodes are placed preferentially at the periphery of the *E. coli* transcription network, thus not perturbing the hierarchical structure shaped by evolution. In yeast, horizontal transfer is generally considered a relatively less important drive for genome evolution.[53] Hence, in both cases we can use our duplication-based analysis to evaluate rewiring in a negative way. In particular, we have considered an observable that quantifies the strength of asymmetric regulations between homology classes, and thus how much of the existing feedback properties can be ascribed to rewiring *versus* duplication. While for the bacterial networks we have found that the empirical value of this parameter corresponds to the null model, in yeast we find a statistically significant trend for the feedback core to come from duplication-divergence, rather than rewiring of existing interactions.

Regarding our findings on horizontal transfers in the *E. coli* transcription network, we should also mention that there is an open debate on this issue. The results are confirmed by an independent study[51] using different methods, and at odds with another parallel study,[54] according to which most paralogs have arisen by horizontal transfer rather than duplication, and thus a duplication-based analysis following the methods of ref. 14 would be ill-defined. While the question remains open, we can state that, independently from the underlying model, in our analysis we found horizontally transferred nodes to be excluded from homology classes, rather than more abundant than expected from a null model.[19]

Finally, we find a strong distinct trend in *E. coli* and yeast, concerning the use and evolution of homodimers *versus* heterodimer transcription factors. In *S. cerevisiae* we found strong indication that heterodimers TFs stem from duplications of ancestral dimeric TFs. The same process seems to be forbidden in *E. coli*, possibly because of the same evolutionary trend erasing crosstalks from duplicate ARs.[19] One can speculate that this ability to make use of heterodimeric binding boosts the combinatorial capacity of a promoter signal integration function.[55]

In conclusion, we presented a comparative analysis of large-scale topological and evolutionary properties of transcription networks in three species, focusing on the global aspects of feedback and hierarchy in regulatory pathways. This analysis confirms that gene duplication is an important drive for the shaping of transcription networks, which follows distinct directions between bacteria, with simple hierarchical networks, and yeast, where more intricate pathways arise. Overall, it appears that yeast is able to sustain a higher complexity in its topological structure, including more feedback and longer pathways, and to explore more freely the possible regulatory interactions stemming from gene duplication, such as feedback produced by duplicate self-regulating or dimeric transcription factors.

## Methods

### Graph growth model

A simple model of network evolution through duplication-divergence was considered. At each time step all the nodes of the graph are duplicated, while the number of edges rises fourfold. This happens for the following reason: for each edge connecting two original (old) nodes (*old–old* edge), duplication of interaction gives rise to edges between the two old nodes and the two duplicate nodes. The original *old–old* edge therefore generates the four *old–old*, *old–new*, *new–old*, *new–new* edges. Duplication of the graph is followed by erasing of edges with prescribed probabilities.[19,31] One can formulate the model with partial or global duplications, and including or not the duplication and removal of self-edges (in this case, we considered the probability of retaining a self-edge equal to that of any other edge). The behavior of this model was compared, through a set of observables, with the observed trends of the experimental data.

### Data sets

We considered the following data sets for the transcription networks. For *E. coli*, the Shen-orr data-set and the larger and more recent RegulonDB5.5.[8,9,56] For *B. subtilis*, DBTBS.[57] For *S. cerevisiae*, the Guelzim[30] data-set and the more recent Balaji[27] data set. The Balaji data set was modified to include auto regulating interactions taken from the literature. In the case of yeast, probably due to the much larger size of the more recent data, there is no compatibility between the two sets of data. Domain architecture data are taken from the SUPERFAMILY database,[58,59] versions 1.61 and 1.69, as in the data sets in ref. 7. Homodimers and heterodimers for *S. cerevisiae*

and *E. coli*, respectively, were obtained from the SGD database[60] and from ref. 61.

### Evaluation of feedback and hierarchy

We used the leaf-removal algorithm[19] on the data-sets (including ARs) and their randomized counterparts. This algorithm prunes the input and output tree-like components of a directed network leaving with a "feedback core" of nodes, where each node is involved in at least one feedback loop. Each iteration of this pruning algorithm defines a hierarchical layer. The main observables we considered were the size of the core and the number of iterations to reach the core. For the evaluation of shortest paths, we used the Dijkstra algorithm,[29] considering the distribution of shortest-path lengths and the longest paths. The results for the empirical networks were compared to randomizations obtained using a standard Markov chain Monte Carlo (MCMC) algorithm[62] or an importance sampling Monte Carlo algorithm[21] that preserve the degree sequence (marginals of the adjacency matrix). We considered random counterparts of the networks where the only constraints come from the in- and out- degree sequences. In particular, we chose not to conserve the number of self-regulators. Finally, to evaluate robustness to noise in the data, we performed the same analyses upon substitution of a fraction of interactions in the original data set with uniformly chosen random ones.

### Evaluation of duplications

We used domain architectures from the SUPERFAMILY database to build protein architecture databases, one for each specie. We then constructed classes of homologous genes using similarity criteria between these architectures, as was done in ref. 14. Two genes are considered homologs if they share the same domains in the same order, neglecting domain repeats. For this analysis, proteins coded by the same operon were considered as separate entities. Since the definition of homology is rather arbitrary it is rather natural to test different definition and observe the stability of results. All these results were filtered for consistency using stricter or looser homology criteria.[37] In the case of the *S. cerevisiae*, and only with respect to the network degree distributions (as shown in Fig. 2), we also considered the notion of homology descending from the gene pairs defined by blocks of doubly conserved syntheny described in ref. 38.

### Duplication of auto regulators

We assessed the evolution of the auto regulated transcription factors by studying the distribution of the subgraphs shown in Fig. 3 within homology classes. Quantitative measure was given by the evaluation of observable $a$, as described in the main text, which was computed on both the experimental data set and on the randomized instances of the null model described later.

### Interaction and rewiring between TFs

We studied the distribution of regulating interaction within and across homology classes. A number of observables were implemented to describe the relationship standing between the architecture data set and the transcription network data set. To quantify the effect of rewiring we introduced the observable $R$ as described in main text. Again, this observable was evaluated upon the experimental data set and upon the null model.

### Evaluation of coevolution of dimer TFs

We studied the evolutionary properties of homodimers and heterodimers with respect to our duplication and divergence model. We assessed the co-existence of homo- and heterodimers within the homology classes by evaluating the sum (over all the homology classes) of the number of all dimeric pairs found in each homology class. The same analysis was performed upon each randomized instance of the null model.

### Null model for duplication

Most of the measures performed upon the experimental data sets were compared with a null model of homology. Keeping fixed the homology classes, we randomly shuffled the architecture associations to the gene names. This randomizes the correlations existing between homology properties (which are responsible of class generation) and the other data sets (adjacency list of transcription network, homodimers, heterodimer pairs). The advantage of this null model lies in the fact that no experimental database is actually randomized, only their interactions. This is important as this randomization does not destroy the homology information nor the global (large scale) properties of the network. Finally, gene name shuffling was done separately for TFs and TGs, as was done in ref. 14, due to their inherently different DNA-binding properties (which depend on their domains). The data shown in this paper correspond to $10^5$ randomized instances of the null model, allowing us to estimate $P$-values larger than $1 \times 10^{-5}$.

## Acknowledgements

We would like to thank Kirill Evlampiev for useful discussions and work on the asymptotics of the theoretical model, Diana Fusco for help with the subgraph analysis, Emmanuel Levy for providing us with a list of *E. coli* homodimer TFs from 3D complex, and M. Madan Babu for critical reading of this manuscript.

## References

1 F. D. Russo and T. J. Silhavy, *Trends Microbiol.*, 1993, **1**(8), 306–10.
2 A. L. Perraud, V. Weiss and R. Gross, *Trends Microbiol.*, 1999, **7**(3), 115–20.
3 E. Perez-Rueda and J. Collado-Vides, *Nucleic Acids Res.*, 2000, **28**(8), 1838–47.
4 D. F. Browning and S. J. Busby, *Nat. Rev. Microbiol.*, 2004, **2**(1), 57–65.
5 L. E. Ulrich, E. V. Koonin and I. B. Zhulin, *Trends Microbiol.*, 2005, **13**(2), 52–6.
6 S. Balaji, M. M. Babu and L. Aravind, *J. Mol. Biol.*, 2007, **372**(4), 1108–22.
7 M. Babu, N. Luscombe, L. Aravind, M. Gerstein and S. Teichmann, *Curr. Opin. Struct. Biol.*, 2004, **14**(3), 283–91.
8 S. Shen-Orr, R. Milo, S. Mangan and U. Alon, *Nat. Genet.*, 2002, **31**(1), 64–8.

 

9 H. Salgado, A. Santos-Zavaleta, S. Gama-Castro, M. Peralta-Gil, M. I. Penaloza-Spinola, A. Martinez-Antonio, P. D. Karp and J. Collado-Vides, *BMC Bioinformatics*, 2006, **7**(1), 5.
10 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**(5594), 799–804.
11 C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J. B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok, M. Kellis, P. A. Rolfe, K. T. Takusagawa, E. S. Lander, D. K. Gifford, E. Fraenkel and R. A. Young, *Nature*, 2004, **431**(7004), 99–104.
12 R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer and U. Alon, *Science*, 2004, **303**(5663), 1538–42.
13 P. Warren and P. ten Wolde, *J. Mol. Biol.*, 2004, **342**(5), 1379–90.
14 S. A. Teichmann and M. M. Babu, *Nat. Genet.*, 2004, **36**(5), 492–6.
15 H. Ma, J. Buer and A. Zeng, *BMC Bioinformatics*, 2004, **5**, 199.
16 H. Ma, B. Kumar, U. Ditges, F. Gunzer, J. Buer and A. Zeng, *Nucleic Acids Res.*, 2004, **32**(22), 6643–9.
17 H. Yu and M. Gerstein, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**(40), 14724–31.
18 G. Balazsi, A. L. Barabasi and Z. N. Oltvai, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**(22), 7841–6.
19 M. Cosentino Lagomarsino, P. Jona, B. Bassetti and H. Isambert, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**(13), 5516–20.
20 R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii and U. Alon, *Science*, 2002, **298**(5594), 824–7.
21 D. Fusco, B. Bassetti, P. Jona and M. Cosentino Lagomarsino, *Bioinformatics*, 2007, **23**(24), 3388–90.
22 D. Thieffry, A. Huerta, E. Perez-Rueda and J. Collado-Vides, *Bioessays*, 1998, **20**(5), 433–40.
23 M. S. Gelfand, *Curr. Opin. Struct. Biol.*, 2006, **16**(3), 420–9.
24 G. Conant and A. Wagner, *Nat. Genet.*, 2003, **34**(3), 264–6.
25 E. Dekel, S. Mangan and U. Alon, *Phys. Biol.*, 2005, **2**(1–2), 81–8.
26 A. Mazurie, S. Bottani and M. Vergassola, *Genome Biol.*, 2005, **6**(4), R35.
27 S. Balaji, M. M. Babu, L. M. Iyer, N. M. Luscombe and L. Aravind, *J. Mol. Biol.*, 2006, **360**(1), 213–27.
28 M. Cosentino Lagomarsino, B. Bassetti and P. Jona, *Lecture Notes in Computer Science*, Springer, Berlin/Heidelberg, 2006, vol. 4210, pp. 227–241.
29 E. W. Dijkstra, *Numerische Mathematik*, 1959, **1**, 269–271.
30 N. Guelzim, S. Bottani, P. Bourgine and F. Kepes, *Nat. Genet.*, 2002, **31**(1), 60–3.
31 K. Evlampiev, *Modélisation de réseaux biologiques*, PhD thesis, Univ. Paris VI/Curie Institute, 2007.
32 K. Evlampiev and H. Isambert, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**(29), 9863–9868.
33 K. Evlampiev and H. Isambert, *BMC Syst. Biol.*, 2007, **1**(1), 49.
34 H. Isambert, *Communicative and Integrative Biology*, 2008.
35 M. Madan Babu and S. A. Teichmann, *Nucleic Acids Res.*, 2003, **31**(4), 1234–44.
36 P. Bork and R. F. Doolittle, *Proc. Natl. Acad. Sci. U. S. A.*, 1992, **89**(19), 8990–4.
37 D. Fusco, F. Grassi, A. L. Sellerio, D. Corà, B. Bassetti, M. Caselle and M. Cosentino Lagomarsino, submitted.
38 M. Kellis, B. W. Birren and E. S. Lander, *Nature*, 2004, **428**(6983), 617–24.
39 I. Ispolatov, A. Yuryev, I. Mazo and S. Maslov, *Nucleic Acids Res.*, 2005, **33**(11), 3629–35.
40 E. D. Levy, J. B. Pereira-Leal, C. Chothia and S. A. Teichmann, *PLoS Comput. Biol.*, 2006, **2**(11), e155.
41 J. B. Pereira-Leal, E. D. Levy, C. Kamp and S. A. Teichmann, *Genome Biol.*, 2007, **8**(4), R51.
42 J. Jeong and P. Berman, *BMC Syst. Biol.*, 2008, **2**, 12.
43 N. Rosenfeld, M. Elowitz and U. Alon, *J. Mol. Biol.*, 2002, **323**(5), 785–93.
44 M. Isalan, C. Lemerle, K. Michalodimitrakis, C. Horn, P. Beltrao, E. Raineri, M. Garriga-Canut and L. Serrano, *Nature*, 2008, **452**(7189), 840–5.
45 I. Wapinski, A. Pfeffer, N. Friedman and A. Regev, *Nature*, 2007, **449**(7158), 54–61.
46 J. J. Ward and J. M. Thornton, *PLoS Comput. Biol.*, 2007, **3**(10), 1993–2002.
47 F. Poelwijk, D. Kiviet and S. Tans, *PLoS Comput. Biol.*, 2006, **2**(5 e58), 0467.
48 J. Ihmels, S. Bergmann, M. Gerami-Nejad, I. Yanai, M. McClellan, J. Berman and N. Barkai, *Science*, 2005, **309**(5736), 938–40.
49 A. Tanay, A. Regev and R. Shamir, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**(20), 7203–8.
50 B. B. Tuch, D. J. Galgoczy, A. D. Hernday, H. Li and A. D. Johnson, *PLoS Biol.*, 2008, **6**(2), e38.
51 M. J. Lercher and C. Pal, *Mol. Biol. Evol.*, 2008, **25**(3), 559–67.
52 D. A. Rodionov, I. L. Dubchak, A. P. Arkin, E. J. Alm and M. S. Gelfand, *PLoS Comput. Biol.*, 2005, **1**(5), e55.
53 C. Pal, B. Papp and M. Lercher, *Nat. Genet.*, 2005, **37**(12), 1372–5.
54 M. N. Price, P. S. Dehal and A. P. Arkin, *Genome Biol.*, 2008, **9**(1), R4.
55 L. Bintu, N. Buchler, H. Garcia, U. Gerland, T. Hwa, J. Kondev and R. Phillips, *Curr. Opin. Genet. Dev.*, 2005, **15**(2), 116–24.
56 H. Salgado, S. Gama-Castro, M. Peralta-Gil, E. Diaz-Peredo, F. Sanchez-Solano, A. Santos-Zavaleta, I. Martinez-Flores, V. Jimenez-Jacinto, C. Bonavides-Martinez, J. Segura-Salazar, A. Martinez-Antonio and J. Collado-Vides, *Nucleic Acids Res.*, 2006, **34**(Database issue), D394–7.
57 Y. Makita, M. Nakao, N. Ogasawara and K. Nakai, *Nucleic Acids Res.*, 2004, **32**(Database issue), D75–7.
58 J. Gough, K. Karplus, R. Hughey and C. Chothia, *J. Mol. Biol.*, 2001, **313**(4), 903–19.
59 D. Wilson, M. Madera, C. Vogel, C. Chothia and J. Gough, *Nucleic Acids Res.*, 2007, **35**(Database issue), D308–13.
60 K. R. Christie, S. Weng, R. Balakrishnan, M. C. Costanzo, K. Dolinski, S. S. Dwight, S. R. Engel, B. Feierbach, D. G. Fisk, J. E. Hirschman, E. L. Hong, L. Issel-Tarver, R. Nash, A. Sethuraman, B. Starr, C. L. Theesfeld, R. Andrada, G. Binkley, Q. Dong, C. Lane, M. Schroeder, D. Botstein and J. M. Cherry, *Nucleic Acids Res.*, 2004, **32**(Database issue), D311–4.
61 G. Butland, J. M. Peregrin-Alvarez, J. Li, W. Yang, X. Yang, V. Canadien, A. Starostine, D. Richards, B. Beattie, N. Krogan, M. Davey, J. Parkinson, J. Greenblatt and A. Emili, *Nature*, 2005, **433**(7025), 531–7.
62 A. Rao, R. Jana and S. Bandyopadhyay, *Indian J. Stat.*, 1996, **58**(A), 225–242.