Molecular BioSystems

Cite this: *Mol. BioSyst.*, 2012, **8**, 3028–3035

www.rsc.org/molecularbiosystems

PAPER

Extensive cross-talk and global regulators identified from an analysis of the integrated transcriptional and signaling network in *Escherichia coli*†

Lucas Antiqueira,*[a] Sarath Chandra Janga[b] and Luciano da Fontoura Costa[c]

Received 14th July 2012, Accepted 16th August 2012
DOI: 10.1039/c2mb25279a

To understand the regulatory dynamics of transcription factors (TFs) and their interplay with other cellular components we have integrated transcriptional, protein–protein and the allosteric or equivalent interactions which mediate the physiological activity of TFs in *Escherichia coli*. To study this integrated network we computed a set of network measurements followed by principal component analysis (PCA), investigated the correlations between network structure and dynamics, and carried out a procedure for motif detection. In particular, we show that outliers identified in the integrated network based on their network properties correspond to previously characterized global transcriptional regulators. Furthermore, outliers are highly and widely expressed across conditions, thus supporting their global nature in controlling many genes in the cell. Motifs revealed that TFs not only interact physically with each other but also obtain feedback from signals delivered by signaling proteins supporting the extensive cross-talk between different types of networks. Our analysis can lead to the development of a general framework for detecting and understanding global regulatory factors in regulatory networks and reinforces the importance of integrating multiple types of interactions in underpinning the interrelationships between them.

1 Introduction

The field of complex networks provides robust tools that researchers in biology can use to represent, characterize and model several problems of interest.[1,2] This is possible since the mathematical concept of graph can be employed whenever a group of interrelated discrete entities are present. In the context of a cell, many different processes such as those driven by metabolic pathways, protein–protein and transcriptional regulatory interactions can be represented as networks.[3–5] However, while the majority of the studies in this area use these networks in an isolated manner, a more comprehensive understanding of the cell requires an integration of different types of cellular interactions – one of the goals of systems biology. Although developments have been made in this direction,[6–11] most studies are limited to understanding particular sub-systems.

One of the fundamental processes even in a simple unicellular biological system such as bacteria is the process of transcriptional regulation. Recent years have seen abundant information accumulating for transcriptional regulation, which has enabled us to model the resulting interactions as a network of transcriptional interactions in bacteria such as *Escherichia coli* and *Bacillus subtilis*.[12–14] While a number of studies have understood these transcriptional networks, most of them have been limited to modeling them as transcription factors (TFs) controlling a set of target genes (TGs).[5,15–18] An additional limitation to the current studies is the employment of only a small set of relatively simple network measurements (*e.g.* degree distribution and clustering coefficient) in a mutually exclusive manner, to understand the local properties of a node.[19–21] Consequently, questions have been raised on the generality of the trends.[15] Also of relevance in this context is the link between the structure and dynamics of a network which is most often neglected when studying global properties. Understanding this link becomes important as is demonstrated in the case of network synchronization in cortical network analyses.[22,23] Another area of complex networks which is extensively explored is the very organization of biological networks into modules and motifs, following the notion that cellular processes are modular in nature.[24] This has given rise to recurring subgraphs or patterns (motifs) which can perform independent functions. For instance, motifs have

[a] *Institute of Mathematical and Computer Sciences, University of São Paulo, PO Box 668, 13560-970, São Carlos, SP, Brazil. E-mail: lantiq@icmc.usp.br; Fax: +55 16 3373 9650; Tel: +55 16 3373 6638*
[b] *School of Informatics, Indiana University Purdue University, Indianapolis, Indiana, Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, 719 Indiana Ave Ste 319, Walker Plaza Building, Indianapolis, Indiana 46202, USA. E-mail: scjanga@iupui.edu; Tel: +1 3172784147*
[c] *Institute of Physics of São Carlos, University of São Paulo, PO Box 369, 13560-970, São Carlos, SP, Brazil. E-mail: ldfcosta@gmail.com; Fax: +55 16 3373 9879; Tel: +55 16 3373 9858*
† Electronic supplementary information (ESI) available. See DOI: 10.1039/c2mb25279a

been investigated in transcription regulatory networks[5] and protein–protein interaction networks.[25]

In this study, we take an integrated approach to study feedback mechanisms in the transcriptional network of a bacterium by incorporating the different aspects discussed above – namely (i) integration of different kinds of networks, (ii) use of more sophisticated network measurements, (iii) understanding the structure–dynamics link and (iv) network motif discovery. We constructed a unified network formed by transcriptional regulatory interactions,[12] metabolic and signaling feedback[26] and protein–protein interactions[21,27] in the bacterium *E. coli*. The transcriptional regulatory network was taken as the base network to which new edges defined by the metabolic and protein interactions were added, resulting in what we call an *integrated network*. We first analyzed the structural organization of the integrated network as well as of the three individual networks considered, which is done by using traditional network measurements, such as degree, clustering coefficient and length of shortest path, as well as more sophisticated metrics such as hierarchical measurements.[28] This approach allowed a global characterization of the structural properties of these networks, *i.e.* it provided parameters to assess their overall organization. In order to complement the characterization of the integrated network we identified nodes having structural properties deviating from the rest of the network. These structural outliers form a group of uncommon nodes that can then be analyzed according to their biological function. To identify them, we calculated a set of measurements for each node, including local (degree and clustering coefficient) and non-local (betweenness centrality, shortest paths and hierarchically-based) features, which were taken as input to principal component analysis.[29] Furthermore, the integrated network was investigated with respect to dynamics (diffusion). More specifically, we used the random walk model[30] to simulate the interaction between genes in terms of the relative frequency of node activation (called here *activity*). With the purpose of investigating how the structure is related to the activation dynamics we evaluated the correlation between in-/out-degrees and activity, thus allowing the identification of dynamical outliers: a group of uncommon genes that are weakly activated even though they control many other genes. Our analysis revealed that outliers identified in the integrated network are global regulators in the transcriptional regulatory network. In addition, we show that outliers are also significantly highly and widely expressed across conditions therefore supporting their deviation from the general trend at the network level. Finally, we identified motifs of sizes up to four in the integrated network and performed a detailed analysis of the origins of 3-node motifs, *i.e.* we investigated how each of the underlying feedback mechanisms contributed to the formation of these motifs in the integrated network. This motif analysis allowed us to show that there is a dense cross-talk between transcriptional regulation and protein–protein interactions in the cell.

2 Methods

In this work we first generated an integrated network formed by transcriptional, protein–protein and metabolic feedback interactions and then focused on understanding this network from both structural and dynamical perspectives. The first step involved the construction of the network, where we integrated the three aforementioned types of biological interactions. The second step involved a structural outlier investigation that employs a set of network measurements and principal component analysis. The third step involved understanding the link between structure and function of TFs by relating diffusion activity to structural network measurements (namely, in- and out-degrees). The last step comprised the identification of motifs to analyze particularly relevant subgraph patterns in the integrated network. We describe each of these steps in detail in the following sections and depict them as flow charts in Fig. 1–4.

2.1 Data integration

We have employed a transcriptional regulatory network (TRN) as the basis for the integrated network developed in this work (Fig. 1). The TRN has directed edges, encoding transcription factors (TFs) regulating protein coding genes in *E. coli*, and has been obtained from the RegulonDB database.[12] Note that by definition some genes, which do not encode for TFs, do not regulate other genes in such a network and are called target genes (TGs); in contrast, genes regulating other genes are called transcription factors (TFs). For TFs which work as heteromeric dimers we have considered the regulation by both the subunits to simplify the simulations. This network has $N = 1521$ nodes (1352 TGs, 169 TFs) and average in- and out-degrees $\langle k_{in} \rangle = \langle k_{out} \rangle = 2.32$. In- and out-degrees are the number of in- and out-going connections of a node, respectively. Notice that the averages of these measurements are always equal for any directed network.

The TRN was then complemented with edges derived from a metabolic and signaling feedback network (MSFN) of *E. coli* published earlier.[26] This is another directed network, with $N = 437$ nodes and average in- and out-degrees $\langle k_{in} \rangle = \langle k_{out} \rangle = 0.91$. It is essentially comprised of signal genes which have the ability to produce cellular signals (either metabolites

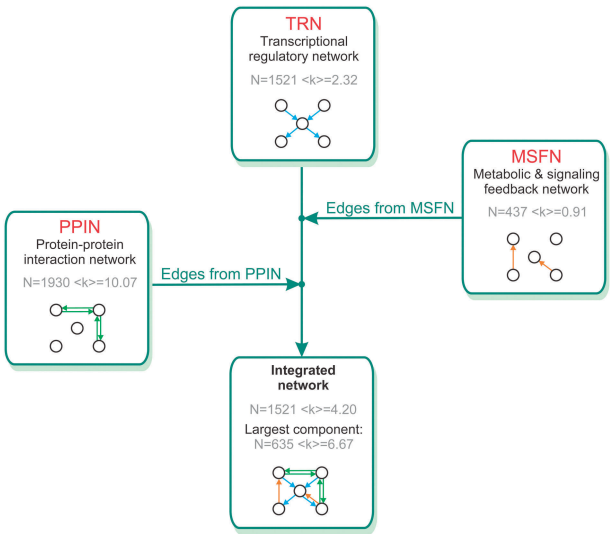

Fig. 1 Flow chart describing integration of the different types of networks (TRN, MSFN, PPIN) to generate the final network used in this study.

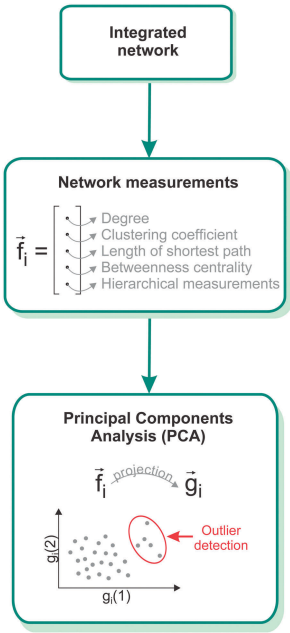

Fig. 2 Flow chart showing the structural outlier detection. Briefly this involved the computation of vectors composed of structural measurements for each node in the integrated network, which were further projected into two dimensions by using PCA.

transported from the exterior of the cell or produced within the cell or metabolites that have the ability to phosphorylate a response regulator) and hence are responsible for modulating the activity of the TFs. Signal genes as defined in this study are enzymes, transporters or histidine kinases which are responsible for producing these signals. Only interactions between nodes already present in the TRN were taken into account in the process of network integration. Therefore, 240 edges from the MSFN were included into the TRN. We then added the protein–protein interaction network (PPIN) of *E. coli* obtained from a recent large-scale experimental screen, having $N = 1930$ and $\langle k \rangle = 10.07$.[27] This is the only undirected structure considered here. Therefore there is no differentiation between in- and out-degrees in this case. In order to make this network directed and therefore to properly merge it with our directed integrated network we took each undirected edge as a symmetric pair of directed ones (Fig. 1). As earlier, only the nodes already present in the TRN were considered when looking for edges in the PPIN. As a result, 2620 PPIN directed edges (or 1310 undirected) were incorporated into the integrated network.

The whole integrated network (TRN + MSFN + PPIN) has $N = 1521$ nodes and average in- and out-degrees $\langle k_{in} \rangle = \langle k_{out} \rangle = 4.20$. Since traditional simple random walks require a network to be connected (see Section 2.3), we used the largest strongly connected component of the integrated network in our simulation experiments. This restriction was applied to other analyses as well (such as for structural outlier identification) to make the object of study uniform throughout experiments and also to allow proper comparisons between different analysis approaches. Furthermore, since other components are too small (at most with three nodes where 97% of them have only one node) the integrated network is heavily fragmented outside the largest component. This network may be complemented in the future when more data becomes available, possibly allowing the growth of the largest component. Henceforth, when we refer to the integrated network we mean its largest component. This final integrated network has average in- and out-degrees $\langle k_{in} \rangle = \langle k_{out} \rangle = 6.67$ and $N = 635$ nodes, of which 525 are TGs and 110 are TFs.

2.2 Structural outlier analysis

When investigating the properties of the integrated network we looked for nodes having structural properties deviating from the rest of the network (*i.e.* structural outliers). The properties of each node are represented in this study by a feature vector composed of F structural measurements. Principal component analysis (PCA),[31] a multivariate method, was chosen to analyze this F-dimensional space (in this work $F = 12$, see Section 3.2). PCA is a common statistical technique that performs a dimensionality reduction through linear combinations that project the original vectors into a new space (see the complete procedure in the ESI†). Since the first dimensions of the projected vectors preserve most of the information (in terms of data dispersion), we only used the two first dimensions of the projected vectors to visually identify outliers, a method later justified by a detailed analysis of the properties of outliers. Since PCA maximizes data dispersion along its first dimensions and completely removes the correlations (redundancy) between features, we are able to detect outliers using fewer dimensions. Fig. 2 illustrates the whole process of structural outlier detection by combining the computation of structural node measurements and PCA. The measurements are: in- and out-degrees, hierarchical in- and out-degrees, in- and out-clustering coefficients, hierarchical in- and out-clustering coefficients, length of shortest path and betweenness centrality. The ESI† contains further details regarding these measurements.

2.3 Structure–dynamics analysis

The method presented in this section concerns the relationship between a dynamical property and structural measurements. The dynamics occurring on a given node is represented by the frequency of visits of a simple random walker and the structural properties correspond to the in- and out-degrees (see the flow chart in Fig. 3). It is important to bear in mind that such random walk dynamics is intrinsically related to diffusion of activations in the network. In other words, the diffusion corresponds to the average of visits to nodes performed by moving agents along a large number of random walk simulations. Therefore, we are interested in relating the diffusion of activations in the network, as modeled by random walks, and the intrinsic properties of nodes. For instance, it is interesting to check if nodes with many connections are more frequently activated or not.[32,33] Formal definitions of random walk and activity can be seen in the ESI.†

We are interested in analyzing the relationships between in- or out-degrees and diffusion of activity (Fig. S1, ESI†). If the structure is well correlated with dynamics, one of these can be obtained from the other with fairly good precision, allowing the prediction of dynamics from structural measurements.

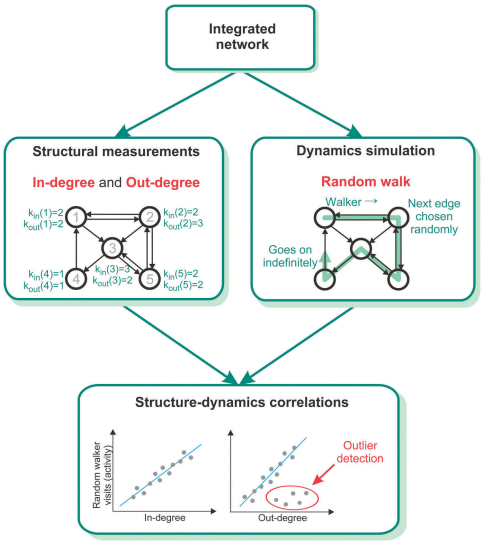

Fig. 3 Flow chart showing the analysis performed to assess structure-dynamics correlations and dynamical outliers. In- and out-degrees correspond to the structure, and the steady-state frequency of visits of a simple random walker represents the dynamics (called activity here).

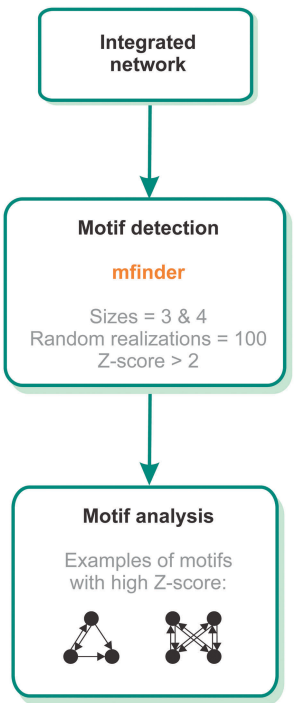

Fig. 4 Motif discovery in the integrated network using the software tool *mfinder*. Motifs with high *Z*-score found in the integrated network are included as examples.

In fact, perfect correlation always occurs in undirected networks, where the activities can be exactly calculated by knowing only the degrees (see Fig. S1A and C, ESI† for example).[33] In the case of directed networks, perfect correlation is implied when the in-degree is equal to the out-degree for every node.[33] Otherwise, in- and out-degrees tend to be uncorrelated or poorly correlated with activity (see Fig. S1B, D and E, ESI†). We use the Pearson correlation coefficient ρ in order to assess the strength of linear correlation,[34] where strong correlations have $|\rho| \to 1$ and weak correlations result in $|\rho| \to 0$. The sign of ρ indicates whether the correlation is positive or negative. In addition to the Pearson coefficient, we also generated the respective scatter-plots in order to provide a complementary means through which weak or medium correlations can be analyzed (as in the examples of Fig. S1, ESI†). In this manner, nodes deviating from the main correlation line can be understood as *dynamical outliers*, thus complementing the structural outlier analysis (Section 2.2).

2.4 Motif analysis

In order to identify connectivity patterns occurring more often than expected by chance in the integrated network we computed motifs of sizes up to four using the motif detection tool *mfinder*[35] (Fig. 4). Full enumeration of subgraphs was chosen as the motif detection method with 100 random network realizations for comparisons.[36] For a given motif its number of occurrences M_{int} in the integrated network was counted, as well as the average number of occurrences μ_{rand} in the randomized counterparts (plus the respective standard deviation σ_{rand}). The *Z*-score of a motif, given by $(M_{\text{int}} - \mu_{\text{rand}})/\sigma_{\text{rand}}$, was used as the main quantifier of motif relevance where only motifs with *Z*-score > 2 were considered. Other parameters were also used to select relevant motifs: *M*-factor > 1.1 and uniqueness ≥ 4, which define, respectively, the minimum fraction $M_{\text{int}}/\mu_{\text{rand}}$ and the minimum number of motif occurrences with different sets of nodes. Besides depicting the general procedure for motif detection, Fig. 4 also illustrates some examples of relevant motifs found in the integrated network.

3 Results

3.1 Analysis of the integrated network

Table S1 (ESI†) presents the averages and standard deviations of the structural measurements calculated for networks TRN, MSFN and PPIN. The measurements are: in- and out-degrees (k_{in} and k_{out}), hierarchical in- and out-degrees at levels 2 and 3 (k^2_{in}, k^3_{out}, k^3_{in} and k^3_{out}), in- and out-clustering coefficients (cc_{in} and cc_{out}), hierarchical in- and out-clustering coefficients at level 2 (cc^2_{in} and cc^2_{out}), length of shortest paths (ℓ) and betweenness centrality (bc) – see the ESI† for definitions. Not surprisingly, the MSFN is the sparsest structure, highly disconnected, with very low degrees and null hierarchical measurements due to the nature of the low-throughput manually curated dataset.[26] On the other hand, the PPIN is the densest network with second and third hierarchies well connected. Nevertheless, shortest paths tend to be longer and centrality values tend to be lower in this network. The TRN has even longer paths, with very low betweenness scores. It is moderately connected along its hierarchies with out-going hierarchies being more sparsely interconnected (see the smaller out-clustering coefficients). Table 1 shows the same measurements computed for the entire integrated network and its largest strongly connected component. Nearly 42% of all nodes were included in the largest component, while 97% of the other components are composed of a single node (the few remaining components have two or three nodes – results not shown).

Table 1 Structural measurements (average and standard deviations) calculated for the integrated network TRN + MSFN + PPIN and its largest strongly connected component

	TRN + MSFN + PPIN	TRN + MSFN + PPIN (largest component)
N	1521	635
TRN edges	3529	1428
MSFN edges	240	205
PPIN edges	2620	2600
k_{in}	4.20 ± 6.05	6.67 ± 8.45
k_{out}	4.20 ± 19.36	6.67 ± 15.31
k^2_{in}	51.52 ± 92.23	98.98 ± 125.28
k^2_{out}	51.52 ± 147.95	98.98 ± 152.01
k^3_{in}	298.18 ± 250.22	424.51 ± 237.06
k^3_{out}	257.49 ± 467.90	415.73 ± 321.56
cc_{in}	0.14 ± 0.18	0.16 ± 0.18
cc_{out}	0.04 ± 0.16	0.10 ± 0.22
cc^2_{in}	0.10 ± 0.13	0.09 ± 0.08
cc^2_{out}	0.02 ± 0.06	0.06 ± 0.09
$\mathscr{C}$	910.54 ± 680.72	3.89 ± 0.84
bc	$8.68 \times 10^{-4} \pm 4.05 \times 10^{-3}$	$4.57 \times 10^{-3} \pm 1.36 \times 10^{-2}$

These numbers indicate that the integrated network is formed by a big component plus a variety of disconnected nodes. This observation is also corroborated by other measurements: degrees at all levels and betweenness centrality are greater for the largest component than in the whole integrated network, while shortest paths are smaller at the largest component. Table 1 also shows how each individual network (TRN, MSFN and PPIN) contributed to the integrated network connectivity. One can notice that almost all MSFN and PPIN edges are preserved in the largest component of the integrated network, whereas more than a half TRN edges are excluded. This fact shows (i) the importance of MSFN and PPIN edges in forming the largest component (*i.e.* interconnecting a considerable share of nodes) and also that (ii) outside the largest component (*i.e.* inside the fragmented portion) almost only TRN edges remain. The later observation indeed supports the notion that there are several peripheral regulatory modules disconnected from the core regulatory network of *E. coli*, either due to the incompleteness of the network or due to their distinct biological roles in contrast to the central metabolism, as has been noted previously.[37]

3.2 Structural and dynamical outliers of the integrated network are composed of global transcriptional regulators

PCA was carried out using feature vectors composed of F = 12 structural measurements: k_{in}, k_{out}, k^2_{in}, k^2_{out}, k^3_{in}, k^3_{out}, cc_{in}, cc_{out}, cc^2_{in}, cc^2_{out}, $\mathscr{C}$ and *bc*, each one calculated individually for every node of the integrated network (only largest component, as explained before). The ESI† contains the specific values of these measurements for each node. Measurements were projected into two-dimensions using PCA, therefore largely preserving the original variation between vectors (89% of total variation) and also making measurements uncorrelated. Fig. 5A shows the projection of all network nodes considering the two first components of PCA. By visually inspecting this plot, we are able to find outliers in the upper part of the figure (gray area in Fig. 5A), *i.e.* nodes whose structure deviates from what is commonly found in the integrated network.

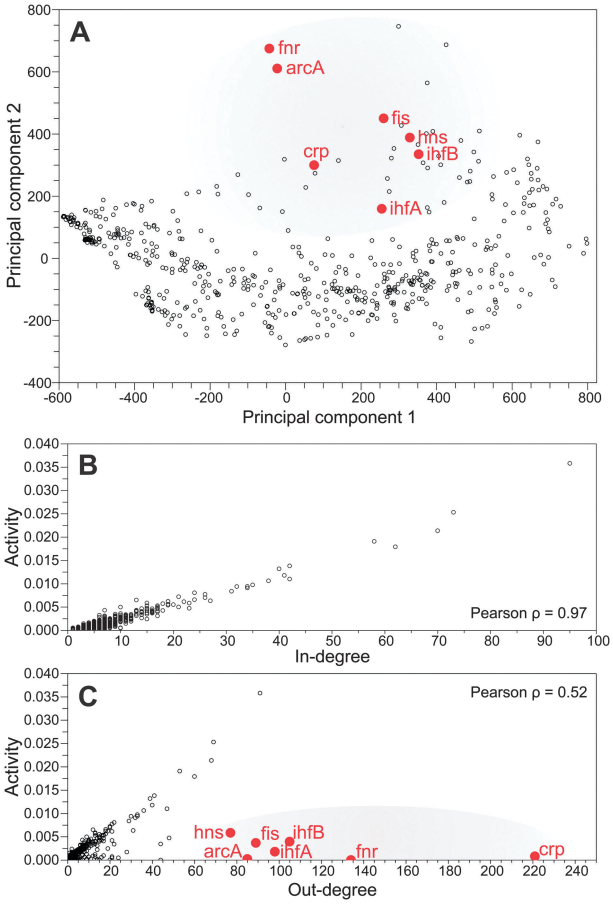

Fig. 5 Structural and dynamical outlier analysis. (A) PCA of the integrated network. Percentage of data variance is 77% in the first axis and 12% in the second. Points falling under the gray area are structural outliers, whereas highlighted points are both structural and dynamical outliers. (B) Correlations in-degree *vs.* activity and (C) out-degree *vs.* activity in the integrated network with dynamical outliers shown under the gray area.

This outlier selection is later justified through a detailed analysis of their network properties – see the remainder of this paragraph. Seven of these structural outliers, highlighted in Fig. 5A (*crp*, *fnr*, *ihfA*, *ihfB*, *fis*, *arcA* and *hns*) are also dynamical outliers (see gray area in Fig. 5C). Note that some structural outliers such as *rplC*, *expB* and *rplV* (they are positioned inside the gray area in Fig. 5A – labels not shown) were not detected as dynamical outliers in Fig. 5C. It is worth noting that each of the highlighted TFs has been described as a global regulator in the transcriptional network of *E. coli* by at least one of the previously published studies.[15,37,38] This observation suggests that dynamical outliers (which are also identified as structural outliers) are very likely to be global transcription factors. Many distinct features distinguish these factors from the rest of the integrated network: (i) all the factors have a much higher out-degree than in-degree, *i.e.* there are many more edges leaving these nodes than coming to them, mostly because of their unusually high out-degree; (ii) their second hierarchical level also presents an unusually high number of out-links (*i.e.* high k^2_{out}) and (iii) $\mathscr{C}$ is smaller for these nodes than for other nodes, *i.e.* these outliers can reach

other nodes by taking only a few steps. Furthermore, we found that outliers can be divided into two groups: (i) *crp*, *ihfA*, *ihfB*, *fis* and *hns*, which present very high betweenness centrality, that is, these nodes take part in a considerable share of shortest paths occurring in the network; and (ii) *fnr* and *arcA*, with uncommonly small k^2_{in}, indicating that their in-going connectivity does not grow as expected when considering a higher hierarchical level. Interestingly, the former is a set of global regulators identified by all of the previous global network analysis surveys.[15,37,38] The latter group is formed by TFs specific to oxygen limitation and/or anaerobic condition that have been found to be more conditionally specific.[39,40] These observations suggest that the structural and dynamical outlier approach presented here can predict global regulators in a given integrated transcriptional network. In particular, we were able to isolate conditionally specific global regulators, a feature which none of the previous methods have been able to achieve.

Dynamical outliers were found in the out-degree *vs.* activity scatter-plot (Fig. 5C). They are nodes which have high out-degree and low activity, therefore departing from the more general correlation occurring among the remaining nodes. Notice that the Pearson correlation coefficient for Fig. 5C is $\rho = 0.52$. When outliers are removed from the scatter-plot the coefficient increases to $\rho = 0.89$, thus corroborating the outlier tendency of departing from the more general linear relationship between out-degree and activity. In biological terms, the activity can be understood as the rate at which each gene is regulated, where the regulatory interactions occur randomly over a known network of possible interactions (*i.e.* simple random walk model). Therefore, these outliers not only have structural features different from the remaining nodes (Fig. 5A) but also present a very odd behavior concerning structure–dynamics correlation. At this point we speculate the main reason for structural outliers being also dynamical outliers: they only receive a few edges at the first hierarchy (and second in some cases) despite being important regulators (*i.e.* with high out-degree) and very close to other nodes (*i.e.* with small paths). Another interesting fact is that many outliers have high betweenness centralities (see the previous paragraph), which means that a great portion of all possible shortest paths includes these nodes. Even being central nodes they fail to be highly active nodes, probably because of their degree imbalance. Finally, in-degree and activity (Fig. 5B) are highly correlated, with Pearson $\rho = 0.97$, therefore presenting no outliers. This also means that in the integrated network the activity of a node can be predicted with a very high confidence while taking only its in-degree into account. On the other hand, outliers in the activity *versus* out-degree plot aid in the identification of global regulators.

3.3 Structural and dynamical outliers in the integrated network are both highly and widely expressed across different experimental conditions

Further analysis of the seven selected outliers (*i.e.* those being simultaneously structural and dynamical outliers) was conducted using data related to the intensity of gene expression in *E. coli* across 302 different conditions available from the M3D database.[41] This procedure allowed us to evaluate outliers not only according to network parameters, but also with respect to their expression context in response to perturbations. In particular we asked whether network outliers are also expression outliers. In order to do so, we summed up normalized expression intensities for each gene considering all expression conditions from the publicly available gene expression atlas and separated results into outliers and non-outliers. The box plot of Fig. 6A shows that outliers are more expressed than most non-outliers (p-value $< 10^{-7}$, Wilcoxon rank sum test). This result builds a strong link between uncommon network features (both structural and dynamical) and expression intensities, therefore supporting a previous observation indicating that the degree of a TF and its expression level are correlated.[17] We also filtered the expression intensity data by defining a threshold above which a given gene expression is present or absent. The threshold is equal to the average expression intensity considering all genes under a given condition. Notice that, though there is one threshold for each condition, we refer to them in the singular for the sake of simplicity. Fig. 6B shows the corresponding boxplot, from which we can observe that the outliers are still more expressed than non-outliers (p-value ~ 0.05, Wilcoxon rank sum test), although with a larger dispersion (see coefficients of variation in the legend of Fig. 6). This expression variance possibly indicates the presence of global regulators which are specific to a handful of conditions. These findings show that the outliers, besides being structurally and dynamically distinct from the other nodes, also correspond to genes with different expression characteristics.

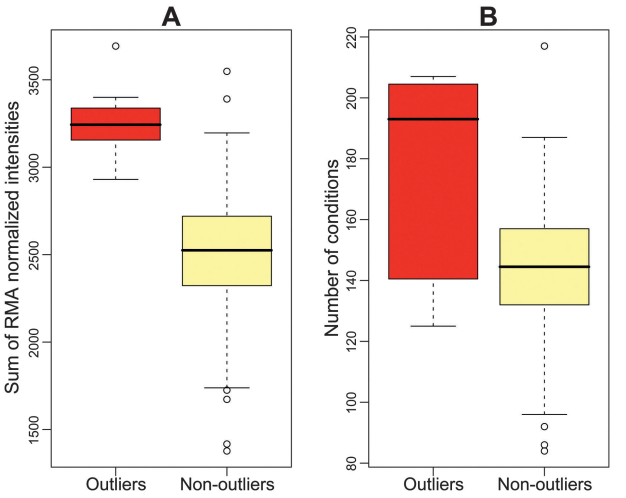

Fig. 6 Box plots of gene expression considering structural and dynamical outliers and the remaining genes, identified here by "non-outliers". (A) refers to the raw expression data, while (B) results from thresholding the gene expression data using the average expression. The ESI† contains the expression intensities used to generate these box plots. Coefficients of variation are: (A) 0.07 for outliers and 0.14 for non-outliers; (B) 0.21 for outliers and 0.15 for non-outliers.

3.4 Integrated network is abundant in novel motif structures which indicate a dense cross-talk between transcriptional regulation and protein complexes

To complement the outlier and expression analyses, we carried out a motif detection procedure on the integrated network – full

motif statistics were included in the ESI.† While outliers are nodes having unusual features, motifs are well-defined interconnected groups of nodes that occur in a network more than expected by chance. Results for 2-node motifs indicated that symmetric links between pairs of nodes frequently occur in the integrated network (with a Z-score = 147.72) because of the inclusion of many symmetric links from the PPIN (see Table 1). Relevant 3-node motifs are depicted in Fig. 7A along with their Z-scores – ESI† also contains the complete list of 3-node motifs. The feed-forward motif[36] appears with Z-score = 2.61 (motif III) and two other motifs with higher Z-scores also occur (notice that these two motifs are in fact the feed-forward with one additional edge leading to the formation of a superposed motif). Fig. S2 (ESI†) shows the integrated network containing only the nodes and edges participating in these 3-node motifs. TRN edges are the most frequent ones with 1041 edges, followed by PPIN edges (713) and MSFN ones (69). Therefore, most part of PPIN and MSFN edges do not take part in 3-node motifs. Moreover, motif types I and III are mainly composed of TRN edges (Fig. 7B), while motif II is strongly based on both TRN and PPI edges. Motif II showed a clear over-representation of instances where transcriptional regulatory and protein–protein interactions were found to be cross-talk. It is noteworthy to mention that in this type of motif the transcriptionally controlled target genes are physically interacting, therefore forming the basis of this motif. Likewise, we also found a significant occurrence of a motif within this type where transcriptional and signaling interactions mutually feedback the target genes controlled by the TFs. Motif III shows the already known feed-forward loop composed entirely of transcription regulations. Additionally, it exhibited a motif structure where a signaling interaction connects the target genes, therefore indicating that the second gene produces a signal which controls the activity of the third gene in the feed-forward loop. These motif instances demonstrate the interplay between metabolic and transcriptional levels *via* the metabolites/signals produced by the signaling genes. Finally, we also found 20 instances of the type I motif where TFs physically interact with each other to control their target gene. Specific instances of 3-node motifs belonging to each of the motif types discussed here are shown in Fig. S3 (ESI†).

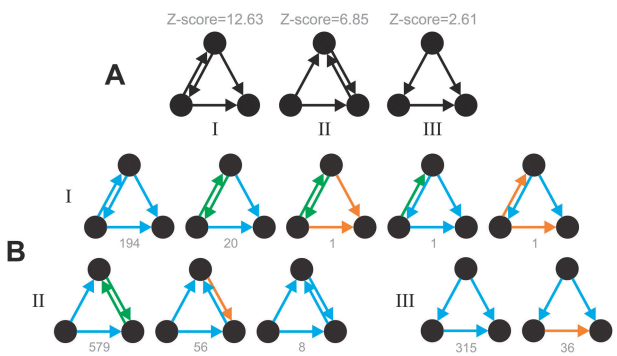

Fig. 7 (A) Three-node motifs and their Z-scores. (B) Three-node motifs separated according to edge type (blue: TRN edge, orange: MSFN edge, green: PPIN edge); the number of occurrences of each motif is also indicated.

Table 2 Ten most frequent nodes/genes in each type of the 3-node motif. For each gene, its absolute (AbsF) and relative (RelF, in %) frequencies of occurrence in each motif type are given. Notice that genes in bold are the outliers of Fig. 5

Motif Type I			Motif Type II			Motif Type III		
Gene	AbsF	RelF	Gene	AbsF	RelF	Gene	AbsF	RelF
ihfB	**92**	**42.4**	**crp**	**170**	**26.4**	**crp**	**115**	**32.8**
ihfA	**91**	**41.9**	**fnr**	**160**	**24.9**	**fnr**	**82**	**23.4**
crp	**47**	**21.7**	**arcA**	**95**	**14.8**	**fis**	**52**	**14.8**
arcA	**45**	**20.7**	aceE	74	11.5	narL	49	14.0
fis	**45**	**20.7**	aceF	53	8.2	fhlA	42	12.0
fnr	**45**	**20.7**	lpd	45	7.0	**ihfA**	**42**	**12.0**
gadX	13	6.0	**ihfA**	**39**	**6.1**	**ihfB**	**42**	**12.0**
hupB	11	5.1	**ihfB**	**39**	**6.1**	pdhR	33	9.4
hns	**10**	**4.6**	**fis**	**37**	**5.8**	fur	19	5.4
gadE	9	4.1	sucC	29	4.5	hyfR	16	4.6

These observations suggest that there is dense networking between transcriptional, physical and signaling interactions of TFs enabling them to integrate diverse cellular processes and stimuli. Motifs with four nodes were also detected in the integrated network (Fig. S4, ESI† also contains the complete list of 4-node motifs). Seven relevant patterns were identified, with three of them having Z-scores much higher than those of the 3-node motifs. Symmetric links occur in almost every 4-node motif, especially in motifs I and V, and most 4-node motifs (except types I and VI) include the feed-forward 3-node motif. Further analysis reveals that outliers (simultaneously structural and dynamical) are also important building blocks of motifs. Table 2 contains the 10 nodes more frequently occurring in 3-node motifs, most of them being the seven outliers *crp*, *fnr*, *ihfA*, *ihfB*, *fis*, *arcA* and *hns*. Outliers also frequently appear in 4-node motifs (Table S2, ESI†), mostly in types II, III and VII. Therefore, structural/dynamical outliers in the integrated network present many distinctive features: (i) they are uncommon nodes with very specific structural and dynamical properties, (ii) they represent genes with different expression characteristics and (iii) they form the foundations of relevant subgraph patterns.

4 Conclusions

Most studies using gene regulatory networks of model organisms have shown the importance of hierarchy and the presence of feed-forward loops. However, there is to our knowledge no study which integrates different processes to unveil the underlying mechanisms controlling the feedback processes of TFs on a global scale. Our observation that there is an extensive cross-talk between TFs and their target genes using protein–protein interactions and signaling interactions suggests that feedback control of TFs is governed by both protein–protein and signaling molecules. Furthermore, the employed method for outlier detection, encompassing structural analysis with dimensionality reduction and structure–dynamics correlations, allowed identification of global regulators. To reinforce the importance of these regulators we showed that they correspond to genes which are both highly and widely expressed across hundreds of conditions, as well as being important building blocks of motifs. We suggest that the analysis employed here can be used as a method to detect global

regulators in regulatory networks of other organisms. To summarize, all these findings illustrate the importance of data integration between different cellular processes.

Acknowledgements

This work was supported by FAPESP – Fundação de Amparo à Pesquisa do Estado de São Paulo (grant numbers 05/00587-5 to L.daF.C. and 06/61743-7 to L.A.) and CNPq – Conselho Nacional de Desenvolvimento Científico e Tecnológico (grant number 301303/06-1 to L.daF.C.).

References

1 A. L. Barabási and Z. N. Oltvai, *Nat. Rev. Genet.*, 2004, **5**, 101–113.
2 L. A. N. Amaral and J. M. Ottino, *Eur. Phys. J. B*, 2004, **38**, 147–162.
3 H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai and A. L. Barabási, *Nature*, 2000, **407**, 651–654.
4 H. Jeong, S. Mason, A. L. Barabási and Z. N. Oltvai, *Nature*, 2001, **411**, 41–42.
5 S. S. Shen-Orr, R. Milo, S. Mangan and U. Alon, *Nat. Genet.*, 2002, **31**, 64–68.
6 P. Hallock and M. A. Thomas, *OMICS*, 2012, **16**, 37–49.
7 H. Lu, B. Shi, G. Wu, Y. Zhang, X. Zhu, Z. Zhang, C. Liu, Y. Zhao, T. Wu, J. Wang and R. Chen, *Biochem. Biophys. Res. Commun.*, 2006, **345**, 302–309.
8 A. Ng, B. Bursteinas, Q. Gao, E. Mollison and M. Zvelebil, *Brief. Bioinf.*, 2006, **7**, 318–330.
9 N. Tenazinha and S. Vinga, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 2011, **8**, 943–958.
10 T. Michoel, A. Joshi, B. Nachtergaele and Y. Van de Peer, *Mol. BioSyst.*, 2011, **7**, 2769–2778.
11 C. Cheng, K.-K. Yan, W. Hwang, J. Qian, N. Bhardwaj, J. Rozowsky, Z. J. Lu, W. Niu, P. Alves, M. Kato, M. Snyder and M. Gerstein, *PLoS Comput. Biol.*, 2011, **7**, e1002190.
12 S. Gama-Castro, V. Jimenez-Jacinto, M. Peralta-Gil, A. Santos-Zavaleta, M. I. Penaloza-Spinola, B. Contreras-Moreira, J. Segura-Salazar, L. Muniz-Rascado, I. Martinez-Flores, H. Salgado, C. Bonavides-Martinez, C. Abreu-Goodger, C. Rodriguez-Penagos, J. Miranda-Rios, E. Morett, E. Merino, A. M. Huerta, L. Trevino-Quintanilla and J. Collado-Vides, *Nucleic Acids Res.*, 2008, **36**, D120–D124.
13 N. Sierro, Y. Makita, M. de Hoon and K. Nakai, *Nucleic Acids Res.*, 2008, **36**, D93–D96.
14 S. C. Janga and J. Collado-Vides, *Res. Microbiol.*, 2007, **158**, 787–794.
15 J. A. Freyre-Gonzalez, J. A. Alonso-Pavon, L. G. Trevino-Quintanilla and J. Collado-Vides, *Genome Biol.*, 2008, **9**, R154.
16 S. C. Janga, H. Salgado, A. Martinez-Antonio and J. Collado-Vides, *Nucleic Acids Res.*, 2007, **35**, 6963–6972.
17 S. C. Janga, H. Salgado and A. Martinez-Antonio, *Nucleic Acids Res.*, 2009, **37**, 3680–3688.
18 M. Madan Babu, S. A. Teichmann and L. Aravind, *J. Mol. Biol.*, 2006, **358**, 614–633.
19 E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai and A. L. Barabási, *Science*, 2002, **297**, 1551–1555.
20 S. Schnell, S. Fortunato and S. Roy, *Proteomics*, 2007, **7**, 961–964.
21 G. Butland, J. M. Peregrín-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**, 531–537.
22 G. Buzsáki, C. Geisler, D. A. Henze and X. J. Wang, *Trends Neurosci.*, 2004, **27**, 186–193.
23 D. Eytan and S. Marom, *J. Neurosci.*, 2006, **26**, 8465–8476.
24 L. H. Hartwell, J. J. Hopfield, S. Leibler and A. W. Murray, *Nature*, 1999, **402**, C47–52.
25 S. Wuchty, Z. N. Oltvai and A. L. Barabási, *Nat. Genet.*, 2003, **35**, 176–179.
26 A. Martinez-Antonio, S. C. Janga, H. Salgado and J. Collado-Vides, *Trends Microbiol.*, 2006, **14**, 22–27.
27 P. Hu, S. C. Janga, M. Babu, J. J. Diaz-Mejia, G. Butland, W. Yang, O. Pogoutse, X. Guo, S. Phanse, P. Wong, S. Chandran, C. Christopoulos, A. Nazarians-Armavil, N. K. Nasseri, G. Musso, M. Ali, N. Nazemof, V. Eroukova, A. Golshani, A. Paccanaro, J. F. Greenblatt, G. Moreno-Hagelsieb and A. Emili, *PLoS Biol.*, 2009, **7**, e96.
28 L. da F. Costa and F. N. Silva, *J. Stat. Phys.*, 2006, **125**, 841–872.
29 L. da F. Costa, F. A. Rodrigues, C. C. Hilgetag and M. Kaiser, *Eur. Lett.*, 2009, **87**, 18008.
30 L. Lovász, in *Combinatorics, Paul Erdős is Eighty*, ed. D. Miklós, V. T. Sós and T. Szőnyi, János Bolyai Mathematical Society, Budapest, 1996, vol. 2, pp. 353–398.
31 R. O. Duda, P. E. Hart and D. G. Stork, *Pattern classification*, John Wiley & Sons, New York, 2000.
32 L. da F. Costa and O. Sporns, *Appl. Phys. Lett.*, 2006, **89**, 013903.
33 L. da F. Costa, O. Sporns, L. Antiqueira, M. G. V. Nunes and O. N. Oliveira Jr., *Appl. Phys. Lett.*, 2007, **91**, 054107.
34 J. L. Myers and A. D. Well, *Research design and statistical analysis*, Lawrence Erlbaum Associates, New Jersey, 2003.
35 N. Kashtan, S. Itzkovitz, R. Milo and U. Alon, 2005, http://www.weizmann.ac.il/mcb/UriAlon/.
36 R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii and U. Alon, *Science*, 2002, **298**, 824–827.
37 A. Martinez-Antonio, S. C. Janga and D. Thieffry, *J. Mol. Biol.*, 2008, **381**, 238–247.
38 A. Martinez-Antonio and J. Collado-Vides, *Curr. Opin. Microbiol.*, 2003, **6**, 482–489.
39 I. Compan and D. Touati, *Mol. Microbiol.*, 1994, **11**, 955–964.
40 S. Spiro and J. R. Guest, *Mol. Microbiol.*, 1988, **2**, 701–707.
41 J. J. Faith, M. E. Driscoll, V. A. Fusaro, E. J. Cosgrove, B. Hayete, F. S. Juhn, S. J. Schneider and T. S. Gardner, *Nucleic Acids Res.*, 2008, **36**, D866–870.