Sanz *et al. BMC Systems Biology* 2012, **6**:110
http://www.biomedcentral.com/1752-0509/6/110

BMC Systems Biology

RESEARCH ARTICLE Open Access

Topological effects of data incompleteness of gene regulatory networks

Joaquin Sanz[1,2], Emanuele Cozzo[1,2], Javier Borge-Holthoefer[1] and Yamir Moreno[1,2*]

Abstract

Background: The topological analysis of biological networks has been a prolific topic in network science during the last decade. A persistent problem with this approach is the inherent uncertainty and noisy nature of the data. One of the cases in which this situation is more marked is that of transcriptional regulatory networks (TRNs) in bacteria. The datasets are incomplete because regulatory pathways associated to a relevant fraction of bacterial genes remain unknown. Furthermore, direction, strengths and signs of the links are sometimes unknown or simply overlooked. Finally, the experimental approaches to infer the regulations are highly heterogeneous, in a way that induces the appearance of systematic experimental-topological correlations. And yet, the quality of the available data increases constantly.

Results: In this work we capitalize on these advances to point out the influence of data (in)completeness and quality on some classical results on topological analysis of TRNs, specially regarding modularity at different levels.

Conclusions: In doing so, we identify the most relevant factors affecting the validity of previous findings, highlighting important caveats to future prokaryotic TRNs topological analysis.

Keywords: Biological networks, Transcriptional regulatory networks, Motifs significance, Community structure, Network superfamilies

Background

As it is commonly noticed in the literature, gene regulation is a complex process involving different phases and biochemical phenomenologies [1,2]. Among these mechanisms, transcriptional control constitutes one of the main resources the cell relies on to respond biochemically to environmental fluctuations and challenges. As a consequence, systematic characterization of TRNs has turned into a subject of high scientific interest [3]. Topological features of TRNs are customarily characterized at all scales using different metrics. At the large scale, genome-wide TRNs are signed and directed networks which present the following features: (i) regulatory proteins –origin of the regulatory interactions of the whole system– represent a small fraction of the total number of nodes; (ii) out-going connectivity patterns are very heterogeneous –a small percentage of global regulators (hubs) send most of the links; and (iii) in-coming link distributions are quite compact: there is a characteristic scale that defines the typical number of regulations each protein receives [4].

Turning to the mesoscale, modularity appears also in TRNs as a key feature to understand the dynamical function of the system. In genome-wide TRNs, each regulator defines its own regulon as the set of nodes directly or indirectly regulated by it. Regulons are then subnetworks, that can be sometimes hierarchically organized; in other occasions, regulons partially overlap in non-trivial ways. Thus, the identification of groups of regulons –or parts of them– interconnected through atypical, dense patterns is expected to store information about the biological role of the proteins within them [5]. The underlying idea is that community structure in biological networks might contribute to unveil functional modularity.

However, perhaps one of the most striking results on topological analysis of TRNs is related to small-scale (sets of 3 or 4 nodes) connectivity patterns, which present

*Correspondence: yamir.moreno@gmail.com
[1]Institute for Biocomputation and Physics of Complex Systems (BIFI), University of Zaragoza, Zaragoza 50009, Spain
[2]Department of Theoretical Physics, University of Zaragoza, Zaragoza 50009, Spain

statistics anything but contingent [6]. Some of these patterns (or *motifs*) have been found to appear much more frequently than expected by random, while others, instead, are underrepresented in real networks. These statistical profiles, measured on different systems, allow the emergence of *network families*, each of which provide a general framework to understand the origin and the dynamical principles of the systems within them [7].

In addition to the aforementioned issues, the experimental challenges underlying the systemic characterization of the TRNs are far from being solved. The quantity and quality of available data on genome-wide transcriptional regulation are significant only for a small set of model organisms. Besides scarcity, the usual problem is related to the heterogeneous quality of the experimental evidences of the regulatory interactions, the building blocks of TRNs. Despite these problems, the amount of high-quality experimental information about transcriptional regulation at systemic level is growing each day, not only within the context of model prokaryotes.

In this work, we analyze three of the best known prokaryotic TRNs, for which these data quality improvements are being more thoroughly incorporated to publicly available data sets. Two of them correspond to the model bacteria *Escherichia coli* [8] and *Bacillus subtilis* [9], while the third one corresponds to the pathogen *Mycobacterium tuberculosis*, whose first network characterizations [10-12] are more recent and incomplete due to the much higher difficulty associated to its wet-lab treatments and protocols. Specifically, the general question we set to answer here is whether robust and biologically relevant conclusions about TRNs can be reached given the current incompleteness of the data, going a step further with respect to other works that had somehow addressed this question previously [13]. Besides, we also show that some topological metrics do depend on the level of detail incorporated in TR maps, in particular, the structure of the mesoscale. Our findings show that extreme care should be taken when strong claims are made based on partial data. This is the case of TRNs superfamilies, which we argue are indeed grouped into a single class.

Results and discussions

Community detection and link attributes

The identification of modules in complex networks has attracted much attention of the scientific community in the last years. A modular view of a network offers a coarse-grained perspective in which nodes are gathered not due to knowledge-based decisions –function, composition, etc.–, but rather on a topological basis –who is connected to whom. To this end Newman put forward the concept of modularity Q [14], which quantifies how far a certain partition is from a random counterpart. From this definition, algorithms and heuristics to optimize modularity (Q) have appeared ever faster and more efficient [15], and generalizations to directed, weighted and signed networks are also available in the literature [16,17]. All these efforts have led to a considerable success regarding the quality of detected community structure in networks, and thus a more complete topological knowledge at this level has been attained. Behind this interest underlies the intuition that the relation between network structure and dynamics is strongly mediated by the mesoscale, and that community structure plays a central role in network formation and functioning. And yet, with few exceptions, link attributes are seldom taken into account.

In this section we intend to underline that interaction direction and sign critically shape the detected community structure of a network. This is ever more dramatic in the case of TRNs, where a sharp distinction must be made between regulators (which mostly emit links) and the rest of the network, which mainly receive them. Also it is peculiar (though not exclusive) of these systems to allow for positive (activating) and negative (inhibitory) relationships. In practice, directions and signs are not always available in the datasets. Regarding directionality, we analyze a system –the TRN of *M. tuberculosis* [12]– for which that is not an actual problem, as regulatory proteins are well identified, i.e. their function as link sources is known. Nevertheless, there are many cases of organisms whose regulatory pathways have not been explicitly identified, and in those cases the real topology is usually replaced by a co-expression network, which acts as an undirected proxy for the true underlying regulatory structure. Unavailability of interaction signs is, on the other hand, a more persistent problem: there exist many experimental approaches to infer a transcriptional regulation that do not inform about the sign of the interaction. Furthermore, there are interaction signs which depend on environmental conditions. Therefore, given the unavoidable incompleteness of the data, we explore whether link attributes determine the network modular structure, and to what extent.

To address the previous question, we perform a systematic comparison of the effects of preserving the original information (sign and direction) in modularity measures and community structure in TRNs. To this end, we will analyze the TRN of *M. tuberculosis* [12], for which we will consider three different topologies: one that preserves all available information (directed-signed, DS); an intermediate one (preserving directions, but not signs –directed-unsigned, DU); and a last one where all fine-grained information is ignored (undirected-unsigned, UU). From the output of this analysis, we provide a way to quantify how much biological information is lost when directions and/or signs are dropped out. Note that the three versions of the network have the same number of nodes N and number of links L, the only differences being those regarding direction and/or the sign of the interactions.

Interaction signs have been compiled from the experimental works enlisted in [12], although signs were not reported there (see [18]).

The modularity expression used hereafter corresponds to its most general definition, i.e. the one that accounts for the existence of directions, weights, signed relations and self-loops, preserving the original information [17]:

$$Q = \frac{w^+}{w^+ + w^-} Q^+ - \frac{w^-}{w^+ + w^-} Q^- \quad (1)$$

This expression generalizes the concept of modularity, and simply computes the contribution to group formation of positive (w^+) and negative (w^-) interactions separately, Q^+ and Q^- respectively, which can be interpreted as the tendency to form communities (positive weights) and that of negative weights to dissolve them. For more detail, Q^+ is defined as

$$Q^+ = \frac{1}{2w^+} \sum_i \sum_j \left(w_{ij}^+ - \frac{w_i^+ w_j^+}{2w^+}) \right) \quad (2)$$

which accounts for the deviation of actual positive weights w_{ij}^+ against a null case random network; the negative counterpart Q^- is defined accordingly, just placing negative weights in the expression. As for our current object of study, links in the network can only take values +1 or -1, and are originally defined as directed.

An intrinsic limitation of modularity maximization, as posed in Eq. 1, is that it provides a single snapshot of the modular structure of the network. However, several topological descriptions of the network coexist at different scales, which is, in general, a fingerprint of complex systems, and particularly relevant in biological ones [19]. A method to overcome this fundamental drawback of typical modularity optimization was put forth in [16]. A parameter r is introduced as a constant self-loop to each node, thus changing the total strength in a network and avoiding the inherent resolution limit of Newman's modularity Q [20]. The shift only affects the property of each node individually and in the same way for all of them. Thus, the original adjacency matrix $\mathbf{A}$ is changed as a function of r: $\mathbf{A_r} = \mathbf{A} + \mathbf{Ir}$. The interesting property of the rescaled topology is that its characteristic scale in terms of modularity has changed. Then the topological structure revealed by optimizing the modularity for $\mathbf{A_r}$ is that of large groups for small values of r, and smaller groups for large values of r, all of which are strictly embedded in the original topology. As an example, the method can uncover each significant resolution level in the well-known synthetic hierarchical network model RB [21], see Figure 1 in [16]. To perform these costly calculations we have used a mixture of heuristics, including extremal optimization and Newman's fast algorithm, as implemented in [22].

Figure 1 (top) represents the number of modules N_c that a combination of Q-maximization heuristics [22] has detected for the three versions of the TRN of *M. tuberculosis*. Each topology has been scrutinized at different scales, screening the parameter r for 200 possible values, in a range such that it yielded an interpretable amount of modules. This range changes for different topologies, thus r is normalized in the plot to allow for comparison. On visual inspection it is apparent that the three topologies present plateaus, where different r values yield similar partitions in terms of N_c. This indicates that certain topological scales are robust and persistent, which might be a clue to identify functionally relevant groups of nodes [16]. Notably, the UU topology presents a single plateau at $N_c = 205$ and then fails to stabilize for larger r's. On the contrary, DU and DS, which retain more information, yield stable partitions at many levels. Although for different r values, these topologies exhibit almost the same behavior regarding plateaus and the number of communities N_c these plateaus present. At this point, one can say that the mesoscale analysis for DU and DS networks allows a richer interpretation in terms of the grouping of nodes, but there is no way to confirm if these are more or less biologically sound, than, for example, the UU topology.

To address this last question, we asked whether the partitions inferred by our method group genes with similar biological functions. The reason underlying this possibility is that genes within a topological community are connected among them by more regulations than average. This fact should imply that they tend to transcript together, as a response to common stimuli and eventually, to perform closely related functions. To do this, we compared the identified communities to the functional classification provided in the Tuberculist database [23]. There are many metrics and indices to compare two clusterings [24-28]. However, we need to rely on an index that does not severely punish different resolution scales: our reference partition categorizes genes in only $N_c^F = 7$ groups, which yields a coarse-grained functional classification of the genome of *M. tuberculosis*. We note that, for the comparison between topological partitions and functional classification, only genes belonging to truly structural functions have been considered. So, "conserved hypotheticals" and "unknown" genes have been excluded, as well as genes with regulatory roles ("regulatory proteins" and "information pathways" genes), which are expected to join transversally the TRN. Any partition with significantly more modules will show low resemblance to the functional one if the index is biased toward literally similar partitions. Thus, we present our results using the Asymmetric Wallace Index AW, which shows the inclusion of a partition into the other. The Asymmetric Wallace Index [29] is the probability that a pair of elements in one cluster of partition A is also in the same cluster of partition B. Let be a clustering A with c_A communities and a clustering B with c_B communities, and let us define

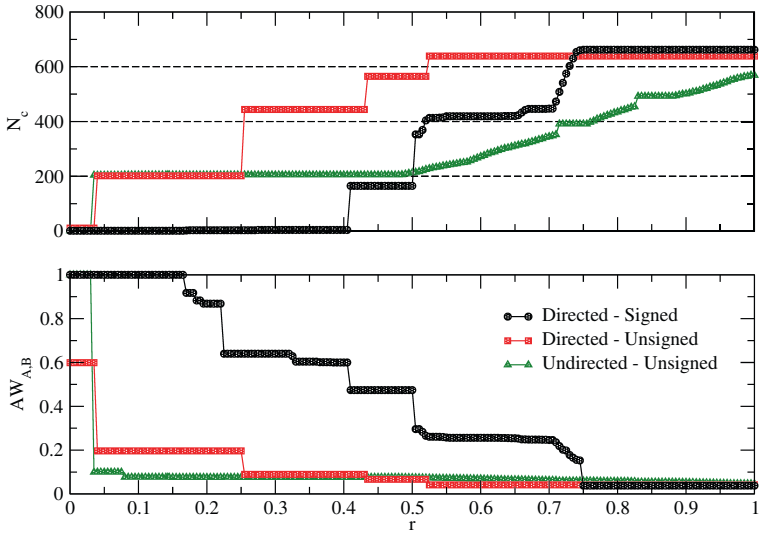

Figure 1 Mesoscale Attributes. Top: number of detected modules N_c as a function of the normalized rescaling parameter r. The mesoscale has been screened only for a range of r yielding an interpretable amount of modules. DU and DS topologies show more than one persistent mesoscopic plateau, whereas the UU topology only has a single plateau made up of around $N_c \approx 200$ communities. Beyond $r = 0.5$, no other stable plateau can be found for this topology. Bottom: the detected community structures for the three versions of the TRN of *M. tuberculosis* are compared to the functional partitions in Tuberculist [23]. DS is the only network that shows significant values of similarity, in terms of the Asymmetric Wallace Index, against the functional partition for a large range of r values.

the confusion matrix M whose rows correspond to the communities of the first clustering (A) and columns correspond to the communities of the second clustering (B). Let the elements of the confusion matrix, $M_{\alpha\beta}$, represent the number of common nodes between community α of the clustering A and community β of the clustering B; the partial sums being $M_{\alpha\cdot} = \sum_\beta M_{\alpha\beta}$ and $M_{\cdot\beta} = \sum_\alpha M_{\alpha\beta}$. Then, $AW_{A,B}$ (how much partition A is embedded in B) is defined as follows:

$$AW_{A,B} = \frac{\sum_{\alpha=1}^{c_A}\sum_{\beta=1}^{c_B} M_{\alpha\beta}(M_{\alpha\beta}-1)}{\sum_{\alpha=1}^{c_A} M_{\alpha\cdot} - 1}. \quad (3)$$

The Asymmetric Wallace index can also be defined the other way around ($AW_{B,A}$), but in this case this is not considered, because detected partitions are systematically more divisive than the functional one, i.e. we are interested in seeing how detected partitions are embedded in the functional one.

Figure 1 (bottom) shows the results for the proposed scheme. Initial results (early r) for the UU and DS networks are artificially high, because $N_c < N_c^F$. Besides this, the plot indicates that only the partitions obtained from the DS topology are significantly similar to the functional one. In fact, beyond the initial stages of the resolution levels, both DU and UU's community structures are far from being embedded in the functional categorization. Quite surprisingly, resolution levels with similar N_c do not entail similar $AW_{A,B}$ values. For instance, the three topologies show at some point a plateau with $N_c \approx 200$. But $AW_{UU,F} \approx 0.1$, $AW_{DU,F} \approx 0.2$ and finally $AW_{DS,F} \approx 0.5$.

These results suggest that the more complete knowledge about link attributes, the richer representation of the mesoscale, in which different levels of topological coarse-graining can be well identified, with possible biodynamical implications that need to be explored.

Motifs significance robustness versus network growth

Exhaustive search of topologically common footprints and systematic differences between different real systems constitutes an important topic in network theory since its very beginning [30]. Along these lines, the classification of networks in families bring light into the evolutionary principles that ultimately yield to the complex topologies that real, evolving systems like TRNs show today [4]. In this sense, the work by Alon and coworkers [7] constitutes a milestone.

In their work, the statistical significance of 3-nodes motifs –triads– was analyzed. The number of appearances of each of the thirteen possible directed structures in real systems was compared to those observed in a null model. The null ensemble was constructed by randomly rewiring the links of the original networks, preserving the number of single links and mutual interactions (as it is done in [7]). The statistical significance of each motif h is then

defined as the Z-score of its number of appearances when compared to the results found in the null ensemble:

$$Z_{\text{score}_h} = \frac{n_h - \langle n_{rand,h} \rangle}{\sigma_{rand,h}} \quad (4)$$

Therefore, computing the Z_{score} for all possible triads in a network yields a 13-dimensional vector that, when normalized, represents the so-called triad significance profile (TSP). From the analysis of different systems' profiles, four superfamilies were identified with common TSPs: two families of non-biological networks –semantic adjacency words maps and social systems– and two families of biological, information processing networks.

Regarding the two biological networks superfamilies originally identified, TRNs of three unicellular organisms were found to conform the first one: yeast, *B.subtilis* and *E.coli*. In Figure 2, panel A, we plot the TSPs that belong to two of the four datasets analyzed by the authors in their original work: yeast [31] and *E.coli* (available at the authors' web site [32]). The second group contains developmental TRNs of eukaryotic cells belonging to pluricellular organisms, signal transduction maps and synaptic networks. In Figure 2, panel B, the TSP of the synaptic wiring map of the nematode *C.elegans* [33] is plotted, as an example of this second superfamily, evidencing the differences with respect to panel A.

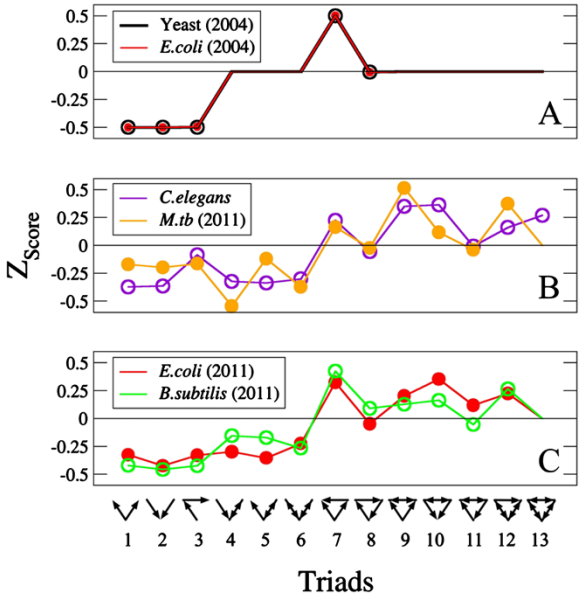

Figure 2 Triad significance profiles (TSPs) of bio-information processing networks. Panel **A**: *E.coli* (2004). TRN of *E.coli* as firstly published in [32,34] ($N = 423$ operons, $L = 519$ regulations plus $SL = 59$ self-regulations). Yeast (2004) TRN of budding yeast [6] ($N = 688$ genes, $L = 1079$ regulations [32,35]). Panel **B**: *C.elegans* neural network [33] ($N = 279$ neurons, $L = 2990$ synapses). *M.tuberculosis* TRN [12] ($N = 1624$ genes, [36] $L = 3169$ regulations plus $SL = 43$ self-regulations). Panel **C**: *E.coli* (2011). Updated TRN of *E.coli* based on RegulonDB, release 7.2, dated on May, 2011 ($N = 1037$ operons, $L = 2574$ regulations plus $SL = 113$ self-regulations). *B.subtilis* (2011), updated TRN of *B.subtilis*, based on DBTBS database [9], (accessed in October, 2011) ($N = 814$ operons, $L = 1294$ regulations plus $SL = 80$ self-regulations).

The biological interpretation of the emergence of the two superfamilies of TRNs –or more generally, bio-information processing networks– proposed in [7] has to do with the typical response times developed by each group of systems. These times are similar to those of single interactions for the networks in the first group (rate-limited networks) but remarkably greater than characteristic interaction times for the systems within the second superfamily (unrate-limited networks).

The recent addition to this scheme of the TRN of *M.tuberculosis* poses an intriguing question. As it is visible to the naked eye in Figure 2 (panel B) its TSP, although belonging to an unicellular organism, has a greater correlation with the representative of the unrate-limited superfamily. The fact that *M.tb.* has these developmental-like topological features at its TRN might be interpreted under a coherent biological picture [12]. The pathogen has an evolutive history tightly bound to its condition of a human intracellular obligate parasite, which could eventually have caused an adaptation of the bacterium to the rhythms and response dynamics of host cells. Indeed, certain stimuli, like hypoxia, yield anomalously slow shifts in *Mycobacterium tuberculosis* gene expression patterns, which can take as much as 80 days until stabilization [11].

The third panel in Figure 2 invalidates the previous hypothesis, and presents the TSPs of the updated TRNs of two bacteria which were initially characterized as rate-limited according to their TSPs. Visible at a glance, the update of the datasets has shifted their TSPs from one superfamily to another, in a way that suggests that the division of the information processing networks into two groups was an effect of data incompleteness.

The key of the change observed in the TSPs stems from the small number of two nodes feedback loops that are observed in unicellular organisms TRNs. Indeed, this possibility was already foreseen in [7] (see footnote 12 there). When feedbacks are absolutely absent from the system under study, as the randomizing algorithm preserves the number of them, feedback loops will also be absent in the null ensemble. This situation makes the Z-scores associated to triads 4, 5, 6, 9, 10, 11, 12 and 13 undefined, as in Figure 2, panel A. As time goes by, such cases have become obsolete: new links have been discovered and added to the growing datasets, and some of them generate feedback loops, which are now present in the triads listed before. In the three updated systems studied, we have found as many as 12 feedback loops in *E.coli* TRN, 9 in *B.subtilis* and 6 in *M.tuberculosis*. The result, after the incorporation of these new feedbacks, suppose that the division between two superfamilies of biological information processing

networks according to their TSPs disappears, affecting the biological interpretation about the eventual relationship between time responses and motifs statistics.

Beyond the discussion on the robustness of motifs statistics that is faced here with a similar spirit of other previous works [13], much has been written about the eventually deep biological implications of anomalous network motifs' statistics as a ubiquitous, topological property of gene regulatory networks. On the one hand, environmental evolutionary adaptation has been claimed to lie underneath this ubiquitous topological treat in gene regulatory circuits [37]. According to this point of view, different environmental requirements could exert different evolutionary pressures to gene expression dynamics which may be correlated to network's topologies at the level of motifs, each of which is believed to offer different dynamical performances, as it has been observed in several precise cases [34,37-39]. Complementarily, recent theoretical studies have addressed how functional, artificial networks required to drive different dynamical functions yield divergent motifs contents [40].

However, as it has been stressed in several works, evolutionary pressures are not the sole mechanism able to generate not-random statistics in networks motifs. Simple models incorporating spatial distribution of nodes [41] or typical mechanisms of network growth assimilable to those which drive gene-regulatory changes upon evolutionary time [42] have been found to generate network motifs without any evolutionary pressure. Under this kind of interpretation, network motifs could appear, not as a consequence of environmental adaptation but rather as a side-effect of some "intrinsic constraints" related to typical mechanisms of genetic material transformation like DNA fragments duplication, deletion, inversion etc [43]. Supporting this hypothesis, a simple but powerful argument is often put forward: topological-bias at the level of TR networks could hardly be a consequence of dynamics-based, natural selection, as in a vast amount of cases transcriptional regulatory mechanisms constitute only one layer of more complex regulatory pathways also coupling translational and post-translational interactions, which are the ultimate responsible of the complex gene expression dynamical patterns observed in the cell [44]. However, comparisons between motifs in gene regulatory networks of different bacteria which should have suffered the effects of entirely comparable "intrinsic constraints" yield slight "fine-tuning" differences in motifs statistics that can be reasonably related to environmental adaptation [12].

The present work does not intend to introduce any additional argument in the debate, which can hardly be considered closed. The reason may be that, as it has been pointed elsewhere, intrinsic constraints and evolutionary pressures are not, definitely, mutually exclusive mechanisms of network transformation [43], and to quantify the relative relevance of each mechanism could result in even a harder task. Our main purpose in this section is, however, to increase our understanding [13] about the robustness of certain topological treats of gene regulatory networks against data incompleteness, as well as to warn about how this analysis affects the network taxonomy scheme proposed in [7].

Systematic correlations between topology and experimental evidence

Experimental techniques used in transcriptional regulation inference are numerous and often subtle [45]. However, usual approaches can be grouped within two main categories. The first approach is based on the explicit detection of the physical protein-DNA interaction between regulators and promoters of target genes. This presents the advantage that only direct operations of regulators on targets can be observed. However, the existence of a protein-DNA interaction under certain in-vitro conditions does not guarantee that it is physiologically relevant in terms of target expression levels.

The alternative approach is essentially based on the generation of mutant strains in which the functionality and/or the expression levels of a certain binding factor are significantly altered with respect to those of the wild type. Then, expression levels of genes which are potentially regulated by the binding factor under study are registered and compared between wild type and mutant strains. In this way, if these different levels of regulator activity yield significantly different target expression measures, one might assume that the regulator is actually acting on the target.

The main advantage of the latter approach is that the sign and strength of the interaction can be determined. However, the analysis cannot distinguish direct regulatory interactions from indirect influences regulator-target mediated by secondary regulatory pathways. Nonetheless, as it can be seen in Table 1, this second kind of methods is responsible for the characterization of an important fraction of the links in our systems. Therefore, a relevant question is whether or not the appearance of indirect, spurious links (as if they were real interactions) might suppose a systematic error responsible of topological bias at a global level.

These hypothetical spurious interactions should appear connecting nodes for which a secondary regulation pathway exists, and its sign should be the same of that secondary route (see Figure 3). So, in our networks, we can identify those "suspicious" links (SLs) connecting nodes for which some secondary via has been already registered, and verify for sign coherence. We will restrict our analysis to those secondary pathways formed by a two-links cascade. The question is how we can know whether this subset of suspicious links presents a higher rate of

Table 1 Well and poorly characterized links

	E.coli	B.subtilis	M.tuberculosis
BA	914	499	726
TELC	1272	856	1290
WCLs	656	323	191
PCLs	1044	262	1344

Number of links reported based upon binding assays (BA) or target expression levels comparison (TELC). Well characterized links (WCLs) are those characterized at least by one methodology of each group (for an exhaustive list of the experimental methods in each group see [18]). Poorly characterized links (PCLs) are reported under methodologies that can not be considered within neither of the two main groups (too generic methods, orthologies based deductions, absence of experimental support etc.). Whilst *B.subtilis* and *E.coli* are relatively well characterized systems, in order to get enough statistics for the *M.tuberculosis* case, we consider identification of consensus sequences as binding proofs.

spurious links than on average. Indeed, among the topologically suspicious links, only those that are characterized by at least one technique of each methodological category –henceforth referred to as well-characterized links (WCLs)–, can safely be considered as non-suspicious direct regulations.

Therefore, the idea is to compare the proportion of well characterized interactions within and outside the subset of topologically suspicious links using Fisher's exact test

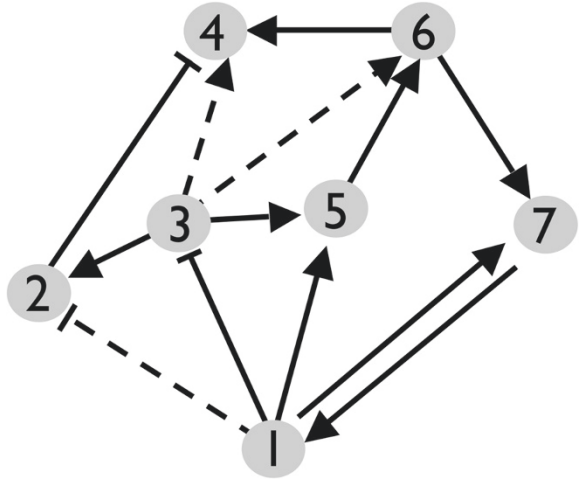

Figure 3 Schematic representation of suspicious links. In the figure, arrows represent activations and right angles inhibitions. Dotted lines correspond to the links that are topologically suspicious, i.e., those for which a secondary regulatory pathway mediated by a third gene is registered and whose sign is coherent. For example, the link connecting nodes 1 and 2 is considered suspicious because of the existence of the pathway 1 to 3 and 3 to 2. The same happens for the link between nodes 3 and 4. In both cases, the condition that the product of the two links making the secondary pathway should coincide with that of the link being considered as suspicious is verified. In this sense, we have also included a case in which the latter condition does not hold: the edge linking node 1 to node 5 is not suspicious because although a two-nodes cascade connecting the same nodes exists, (i.e. 1 to 3 and 3 to 5) it is not sign coherent.

(see Table 2). As it can be seen, suspicious links systematically present a slightly lower proportion of WCLs than non-suspicious links, which could be associated to random fluctuations with respect to the average values with probabilities lower than 2% in each of the systems, being remarkably lower in the case of the TRN of *E.coli*.

This indicates that suspicious links constitute a topologically defined subset of interactions which is systematic and significantly less reliably characterized than on average in all the systems under study. This observation is in agreement with the hypothesis that insufficient experimental methods of transcriptional regulation inference can suppose the systematical observation of topologically-biased spurious links. The problem addressed here seems to critically affect the characterization of the activity of sigma factors. In fact, when we reconstruct the networks under study by considering only transcription factors as regulators and exclude sigma factors, the whole picture significantly changes. Indeed, the percent of suspicious links which are better characterized is even greater than the background, both for *B.subtilis* (45.0% vs 42.2%) and for *E.coli* (46.0% vs 43.1%). For the case of *M.tuberculosis*, the analysis can be hardly conclusive due to the loss of statistics after sigma factors removal (no well characterized link is located within the set of suspicious interactions, now, less than 100 in the whole signed network). These findings, put together, suggest that characterization of sigma factor regulons is more sensitive to the aforementioned issues.

Another issue of interest is whether this experimental bias is topologically relevant. More precisely, we question if this systematic error could quantitatively affect motif statistics in our systems. The key is that these spurious interactions could recurrently transform some motifs into others, and more precisely, focusing on most prominent motifs in number of appearances, this would suppose the systematic, spurious transformation of three-nodes cascades (triad 3) into coherent feedforward loops (triad 7). To test the robustness of the TSPs to the presence of spurious links, we delete in each network a fraction of partially characterized suspicious links up to the point in which the proportion of WCLs among them is comparable to the average background level. This suppose the removal of 324 suspicious links in the TRN of *E.coli*, 59 in the TRN of *B.subtilis* and 114 for the *M. tuberculosis* case. The links to be deleted are randomly chosen within the set of partially characterized suspicious links. Finally, we recalculate the Z_{scores} of all the motifs and compare TSPs with their original values. The results of this process are shown in Table 3, where it can be seen that the statistical significance of cascades and feedforward loops are systematically affected. The interesting fact is that, after the correction, cascades are yet significantly underrepresented while feedforward loops as a whole (i.e.

Table 2 Statistics of suspicious links

	E.coli		B.subtilis		M.tuberculosis	
	WCLs	No WCLs	WCLs	No WCLs	WCLs	No WCLs
SLs	104	628	43	188	11	252
No SLs	552	1285	280	774	180	2141
	H_o p value $< 10^{-17}$		H_o p value < 0.007		H_o p value < 0.019	

Null hypothesis H_o assumes that the proportion of WCLs is the same for SLs or no-SLs. Only signed links are considered.

independently of the signs) continue to appear much more frequently than expected by random. Obviously, the Z_{score} associated to the other triads also varies. But the striking point is that, after normalization, in all the systems the effect of the correction on the TSPs are very limited, as we can see in Figure 4. So, the conclusion is that this kind of systematic error, although modifying the absolute values of motifs' Z_{scores}, does not affect their ratios that are recovered after normalizing the TSPs.

Conclusions

As we have shown here, sources of unreliability can be of diverse nature: from the often unjustified lack of details in link attributes to the lack of key interactions, whose inclusion radically modify motifs' TSPs. As a matter of fact, our first finding convincingly shows that data incompleteness could exert a relevant influence on the topological characterization of the mesoscale in prokaryotic TRNs. More precisely, we have shown how a complete knowledge of link attributes (directions and signs) can yield richer mesoscale structures in TRNs. Secondly, we have also shown that a mere updating of the interactions that make up a TRN in which key regulatory interactions are incorporated, radically modifies previous results based on the analysis of motifs appearances. In fact, some of the previous conclusions do not hold anymore. We have observed that prokaryotic TRNs show motifs significance profiles very similar to those belonging to multicellular, developmental TRNs, signal transduction and neural systems. Finally, experimental mischaracterization of the links has also been studied, and yet, we have found that its influence on motifs statistics is reduced. These results suggest that the evolutionary interplay between topology and dynamics is more similar between regulatory systems of multicellular and unicellular organisms than expected.

Transcriptional Regulatory Networks have been increasingly studied during the last several years. Nowadays, however, their characterization can only be considered provisional, as they consist of incomplete annotations of often heterogeneous and unreliable experimental evidences, computational inferences and theoretical predictions. While working with still incomplete networks could be of valuable help to uncover unknown biochemical pathways, there are situations in which reliable conclusions cannot be obtained. Moreover, we don't even know when the latter is the case. Accuracy and robustness of the results thus require us to be able to assess what results are dependent on the noisy and

Table 3 Variation in Z_{scores}

	E.coli	B.subtilis	M.tb.
Cascade (original)	-4.7 ± 0.2	-6.9 ± 0.4	-2.2 ± 0.1
Cascade (corrected)	-2.9 ± 0.4	-6.9 ± 0.8	-1.1 ± 0.3
FFL (original)	4.7 ± 0.2	6.9 ± 0.4	2.2 ± 0.1
FFL (corrected)	2.5 ± 0.7	7.0 ± 0.8	1.1 ± 0.3

Changes in the Z_{scores} of cascades and feed-forward loops (FFLs) due to systematic mischaracterization of SIs.

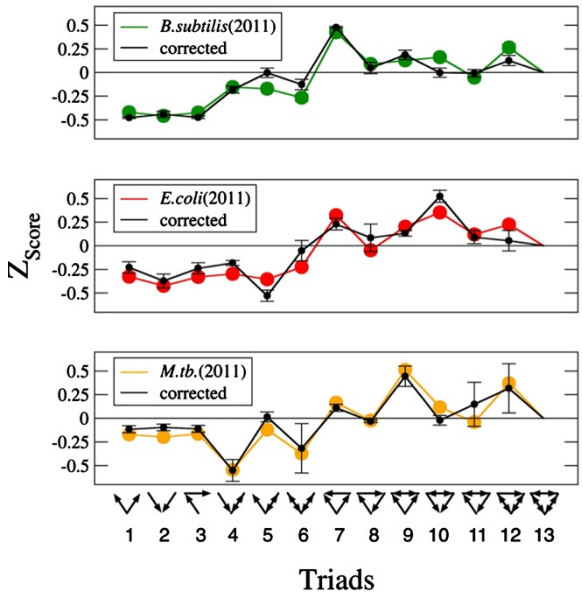

Figure 4 Changes in prokaryotic TSPs due to systematic experimental mischaracterization of links. Original systems present a lower proportion of WCLs than the background in the set of links that connect nodes for which a secondary coherent regulatory pathway has been registered. In corrected networks, the proportion of WCLs is paired to background levels by randomly deleting poorly characterized interactions. This, however, do not affect the TSPs significantly.

uncertain nature of some annotated links. This is crucial if deep biological implications are to be claimed.

Methods

We find the community structure of the networks studied using the modularity concept introduced by Newman [14]. To perform these costly calculations we have used a mixture of heuristics, including extremal optimization and Newmans fast algorithm, as implemented in [22]. On the other hand, the statistical significance of motifs has been calculated as it is customarily done [6,7]. Finally, for an exhaustive list of the experimental methods that have been categorized in different groups, see http://cosnet.bifi.es/researchlines/systems-biology/data.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JS and YM designed research, JS, EC, JBH performed research, JS, EC, JBH and YM analyzed the data, JBH and YM wrote the manuscript. All authors have read and approved the final version.

Acknowledgements

We would like to thank José Alberto Carrodeguas and Alejandra Nelo for helpful comments. This work has been partially supported by MICINN through Grants FIS2008-01240, FIS2009-13364-C02-01, and FIS2011-25167 and by Comunidad de Aragón (Spain) through a grant to FENOL group.

Received: 18 May 2012 Accepted: 6 July 2012
Published: 25 August 2012

References

1. Sorek R, Cossart P: **Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity.** *Nat Rev Genet* 2010, **11:**9–16.
2. Day DA, Tuite MF: **Post-transcriptional gene regulatory mechanisms in eukaryotes: an overview.** *J Endocrinol* 1998, **157:**361–371.
3. Sirbu A, Ruskin H, Crane M: **Comparison of evolutionary algorithms in gene regulatory network model inference.** *BMC Bioinformatics* 2010, **11:**59.
4. Babu M, Teichmann S, Aravind L: **Evolutionary Dynamics of Prokaryotic Transcriptional Regulatory Networks.** *J Mol Biol* 2006, **358:**614–633.
5. Bar-Joseph Z, et al.: **Computational discovery of gene modules and regulatory networks.** *Nat Biotechnol* 2003, **21:**1337–1342.
6. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U: **Network Motifs: Simple Building Blocks of Complex Networks.** *Science* 2002, **298:**824–927.
7. Milo R, Itzkovitz S, Kashtan N, Levitt R, Shen-Orr S, Ayzenshtat I, Sheffer M, Alon U: **Superfamilies of evolved and designed networks.** *Science* 2004, **303:**1538–1542.
8. Gama-Castro S, *et al*: **RegulonDB version 7.0: transcriptional regulation of *Escherichia coli* K-12 integrated within genetic sensory response units (Gensor Units).** *Nuc Acids Res* 2010, **39**(Database issue):D98–D105.
9. Sierro N, Makita Y, de Hoon MJL, Nakai K: **DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information.** *Nuc Acids Res* 2008, **36**(Database issue):D93—D96.
10. Jacques PE, Gervais AL, Cantin M, Lucier JF, Dallaire G, Drouin G, Gaudreau L, Goulet J, Brzezinski R: **MtbRegList, a database dedicated to the analysis of transcriptional regulation in Mycobacterium tuberculosis.** *Bioinformatics* 2005, **21:**2563–2565.
11. Balazsi G, Heath A, Shi L, Gennaro M: **The temporal response of the Mycobacterium tuberculosis gene regulatory network during growth arrest.** *Mol Sys Biol* 2008, **4:**225.
12. Sanz J, Navarro J, Arbués J, Martín C, Marijuán P, Moreno Y: **The transcriptional regulatory network of Mycobacterium tuberculosis.** *PLoS One* 2011, **6**(7):e22178.
13. de Silva E, Thorne T, Ingram P, Agrafioti I, Swire J, Wiuf C, Stumpf MPH: **The effects of incomplete protein interaction data on structural and evolutionary inferences.** *BMC Biol* 2006, **4:**39.
14. Newman M, Girvan M: **Finding and evaluating community structure in networks.** *Phys Rev E* 2004, **69:**026113.
15. Fortunato S: **Community detection in graphs.** *Phys Rep* 2010, **486:**75–174.
16. Arenas A, Fernández A, Gómez S: **Analysis of the structure of complex networks at different resolution levels.** *New J Phys* 2008, **10:**053039.
17. Gómez S, Jensen P, Arenas A: **Analysis of community structure in networks of correlated data.** *Phys Rev E* 2009, **80:**016114.
18. **Signed version of the transcriptional regulatory network of *M.tuberculosis* published at [12].** [http://cosnet.bifi.es/research-lines/systems-biology/data]
19. Spirin V, Gelfand M, Mironov A, Mirny L: **A metabolic network in the evolutionary context: Multiscale structure and modularity.** *Proc Nat Acad Sci* 2006, **103**(23):8774–8779.
20. Fortunato S, Barthélemy M: **Resolution limit in community detection.** *Proc Nat Acad Sci* 2007, **104**(1):36–41.
21. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL: **Hierarchical organization of modularity in metabolic networks.** *Science* 2002, **297**(5586):1551–1555.
22. Gómez S, Fernández A, Borge-Holthoefer J, Arenas A: **radatools.php.** [http://deim.urv.cat/~sgomez/]
23. Lew J, Kapopoulou A, Jones L, Cole S: **Tuberculist: 10 years after.** *Tuberculosis Edinb* 2011, **91**(1):1–7.
24. Kuncheva L, Hadjitodorov S: **Using diversity in cluster ensembles.** In *Systems, Man and Cybernetics, IEEE International Conference on Systems, man and Cybernetics. Volume 2.* 2004:1214–1219.
25. Rand WM: **Objective criteria for the evaluation of clustering methods.** *J Am Stat Asoc* 1971, **66**(336):846–850.
26. Hubert L, Arabie P: **Comparing partitions.** *J Classif* 1985, **2**(1):193–218.
27. Fowlkes EB, Mallows CL: **A Method for Comparing Two Hierarchical Clusterings.** *J Am Stat Asoc* 1983, **78**(383):553–569.
28. Meila M: **Comparing clusterings: an information based distance.** *J Multivariate Anal* 2007, **98**(5):873–895.
29. Wallace D: **A Method for Comparing Two Hierarchical Clusterings: Comment.** *J Am Stat Asoc*, **78**(383):569–576.
30. da Costa LF, Rodrigues FA, Travieso G, Villas-Boas PR: **Characterization of complex networks: A survey of measurements.** *Adv Phy* 2007, **56**(1):167–242.
31. Costanzo M, *et al*: **YPD, PombePD and WormPD: model organism volumes of the BioKnowledge Library, an integrated resource for protein information.** *Nuc Acids Res* 2001, **29**(1):75–79.
32. **Uri Alon's lab website.** [http://www.weizmann.ac.il/mcb/UriAlon/]
33. **Database of synaptic connectivity of C. elegans for computation.** *Technical report of Cybernetic Caenorhabditis elegans Program* 2003. [http://ims.dse.ibaraki.ac.jp/ccep/]
34. Mangan S, Alon U: **Structure and function of the feed-forward loop network motif.** *Proc Nat Acad Sci* 2003, **100**(21):11980–11985.
35. **In [7] –see note 12 there–, feedback loops are cancelled when supposing less than 0.1% of network links. According to that convention, the only feedback loop in yeast TRN –which obviously could not be rewired– is cancelled.**
36. **An operon based representation is not available for the TRN of *Mycobacterium tuberculosis* because of that a global enough experimental characterization of its operon map has not been accomplished yet. To our knowledge, most relevant works in this area –see, for example: Roback P , Beard J , Baumann D , Gille C , Henry K. 2007 A predicted operon map for Mycobacterium tuberculosis.** *Nuc Acid Res*, **35**(15):5085–5095 doi:10.1093/nar/gkm518 – consist yet of general computational predictive tools.
37. Alon U: **Network motifs: theory and experimental approaches.** *Nat Rev Genet* 2007, **8:**450–461.
38. Mangan S, Itzkovitz S, Zaslaver A, Alon U: **The incoherent feed-forward loop accelerates the response-time of the gal system of Escherichia coli.** *J Mol Biol* 2006, **356**(5):1073–1081.

39. Zaslaver A, Mayo AE, Rosemberg R, Bashkin P, Sberro H, Tsalyouk M, Surrette MG, Alon U: **Just-in-time transcription program in metabolic pathways.** *Nat Gen* 2004, **36**(5):486–491.
40. Burda Z, Krzywicki A, Martin OC, Zagorski M: **Motifs emerge from function in model gene regulatory networks.** *Proc Nat Acad Sci* 2011, **108**(42):17263–17268.
41. Artzy-Randrup Y, Fleishman SJ, Ben-Tal N, Stone L: **Comment on Network Motifs: Simple Building Blocks of Complex Networks and Superfamilies of Evolved and Designed Networks.** *Science* 2004, **305:**1107.
42. Dwight Kuo, P, Banzhaf W, Leier A: **Network topology and the evolution of dynamics in an artificial genetic regulatory network model created by whole genome duplication and divergence.** *Biosystems* 2006, **85**(3):177–200.
43. Huang S: **Back to the biology in systems biology: What can we learn from biomolecular networks?** *Briefings Funct Genomics* 2004, **2**(4):279–297.
44. Mazurie A, Bottani S, Vergassola M: **An evolutionary and functional assessment of regulatory network motifs.** *Genome Biol* 2005, **6:**R35.
45. Banerjee N, Zhang M: **Functional genomics as applied to mapping transcription regulatory networks.** *Curr Op Microbiol* 2002, **5:**313–317.

doi:10.1186/1752-0509-6-110
Cite this article as: Sanz *et al.*: Topological effects of data incompleteness of gene regulatory networks. *BMC Systems Biology* 2012 **6**:110.