RESEARCH ARTICLE

Feedbacks from the metabolic network to the genetic network reveal regulatory modules in *E. coli* and *B. subtilis*

Santhust Kumar[1], Saurabh Mahajan[2], Sanjay Jain[1,3]*

1 Department of Physics and Astrophysics, University of Delhi, Delhi 110007, India, **2** National Centre for Biological Sciences, Bangalore, Karnataka 560065, India, **3** Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, United States of America

* jain@physics.du.ac.in

Citation: Kumar S, Mahajan S, Jain S (2018) Feedbacks from the metabolic network to the genetic network reveal regulatory modules in *E. coli* and *B. subtilis*. PLoS ONE 13(10): e0203311. https://doi.org/10.1371/journal.pone.0203311

Editor: Szabolcs Semsey, Niels Bohr Institute, DENMARK

Received: May 5, 2018

Accepted: August 18, 2018

Published: October 4, 2018

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by Department of Biotechnology, Government of India (http://www.dbtindia.nic.in/), grant no. BT/COE/34/SP15246/2015 to SJ; and Science and Engineering Research Board, Government of India (www.serb.gov.in), grant no. SR/S4/MS:791/12 to SJ. SK acknowledges a Junior Research Fellowship and Senior Research Fellowship from the University

Abstract

The genetic regulatory network (GRN) plays a key role in controlling the response of the cell to changes in the environment. Although the structure of GRNs has been the subject of many studies, their large scale structure in the light of feedbacks from the metabolic network (MN) has received relatively little attention. Here we study the causal structure of the GRNs, namely the chain of influence of one component on the other, taking into account feedback from the MN. First we consider the GRNs of *E. coli* and *B. subtilis* without feedback from MN and illustrate their causal structure. Next we augment the GRNs with feedback from their respective MNs by including (a) links from genes coding for enzymes to metabolites produced or consumed in reactions catalyzed by those enzymes and (b) links from metabolites to genes coding for transcription factors whose transcriptional activity the metabolites alter by binding to them. We find that the inclusion of feedback from MN into GRN significantly affects its causal structure, in particular the number of levels and relative positions of nodes in the hierarchy, and the number and size of the strongly connected components (SCCs). We then study the functional significance of the SCCs. For this we identify condition specific feedbacks from the MN into the GRN by retaining only those enzymes that are essential for growth in specific environmental conditions simulated via the technique of flux balance analysis (FBA). We find that the SCCs of the GRN augmented by these feedbacks can be ascribed specific functional roles in the organism. Our algorithmic approach thus reveals relatively autonomous subsystems with specific functionality, or regulatory modules in the organism. This automated approach could be useful in identifying biologically relevant modules in other organisms for which network data is available, but whose biology is less well studied.

1 Introduction

The control of gene expression is central to cellular dynamics. All cellular processes—metabolism, growth, cell division, response to stimuli, etc.—are related to characteristic

Grants Commission, Government of India. SM acknowledges a Senior Research Fellowship from the Department of Biotechnology, Government of India. SJ acknowledges the Addie and Harold Broitman Membership of the Simons Center for Systems Biology, Institute of Advanced Study, Princeton, USA. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

subsets of genes that need to be expressed for the respective processes to take place inside the cell. The expression of genes inside a cell is controlled via a complex process of inter-regulation involving, in part, a network of genes called the gene regulatory network (GRN). Due to the large number of genes and interactions in a GRN and the presence of feedback loops, it becomes difficult to track the causal regulation from one component of the GRN to another. In order to obtain a system level understanding of gene interactions of an organism, an understanding of the structure of its GRN and its design principles is necessary.

Several studies have characterized the GRNs on a global and local scale. The GRNs have been found to follow an exponential in-degree distribution and a power law out-degree distribution [1–3]. They have been shown to possess a hierarchical and modular organization [4–16]. While hierarchy depicts the regulatory flow of information in a system, identifying the modules in a system has also been found useful in comprehending its functional and structural organization [17–20]. Modules have been defined in various ways: as a group of genes that express together [21–24], as a cluster or community of nodes that connect tightly together compared to other nodes in the network [25, 26], as a set of connected nodes revealed upon the removal of common global regulators [27], and as network motifs [28, 29] which have a specific functionality.

Despite its frequent isolated treatment the GRN of an organism functions in conjunction with other networks in the cell, in particular the metabolic network (MN). It has been shown [4] that about half of the transcription factors, which belong to the GRN, bind to small molecules which are part of the MN. In order to have a better understanding of the organisational structure and functioning of the GRN, consideration of the influence of the MN over it is crucial. Several works study the GRN and MN in an integrated manner to explore various structural, dynamical or evolutionary aspects of their interplay [11, 30–35]. Recently a combined network of GRN, MN and protein interactions has been used to explore the cascading impact of perturbations originating in different parts of the combined network [36]. However, the global architecture of the integrated GRN and MN remains to be elucidated. Here we study some aspects of this architecture, in particular focusing on the effect of feedback from MN to GRN on the hierarchical and modular structure involved in the overall control of metabolism.

To this end, we obtain the data about the genetic regulatory interactions and metabolism of two microorganisms, *E. coli* and *B. subtilis*, from a number of sources in literature [37–43]. Certain metabolites bind to transcription factors (TFs) and alter their gene regulatory activity. We add to the GRN nodes corresponding to such metabolites as well as links from these nodes to the genes coding for TFs to which the metabolites bind. In addition we include links to these metabolites from genes coding for enzymes which catalyze the reactions of these metabolites. The augmented GRNs so obtained consist of approximately 3300 nodes and 9300 edges (*E. coli*), and 1700 nodes and 3500 edges (*B. subtilis*).

We first organize each GRN (without including the metabolite nodes) into a hierarchical structure by using a modified form of the vertex sort algorithm of Jothi et al [13] which involves finding strongly connected components (SCCs) of the network, thereby elucidating its regions of feedback and causal structure.

Second, we augment the GRN by introducing nodes and links corresponding to metabolites belonging to the MN as described above, and discuss how the hierarchical structure changes. We show that GRNs retain their characteristic hierarchical structure upon including the feedbacks from metabolism; however they become more complicated and the causal ordering governed by the relative positions of genes among the various levels in the hierarchy is significantly altered.

Third, we exploit the fact that not all feedbacks from the MN are functional at a given time. To this end we simplify the augmented network by first identifying functionally relevant feedbacks from the MN into the GRN under a number of simulated environmental conditions (ECs) using the technique of flux balance analysis (FBA) [44], and then augmenting the GRN with only these functionally relevant feedbacks. The structure of the augmented network thus obtained is easier to interpret in terms of dynamics and biological function. We then interpret the resulting SCCs of the network as modules, following Axelsen et al [10] and Rodriguez-Caso et al [14]. Our definition of modules differs somewhat from that of [10, 14] in that we do not limit the module to just the SCC, but also consider the proximal regulatory circuit around the nodes belonging to the SCC. This in conjunction with the fact that metabolite nodes are also part of our network allows us to assign a specific functional role to most of the SCCs. This role follows from the circuit diagram of the SCC and the manner in which it is embedded in the full network, which makes its qualitative dynamics and biological function quite evident. This effectively provides a list of biologically relevant dynamical sub-systems of the cell for further investigation. In addition to finding and classifying the important modules of the joint GRN and MN of *E. coli* and *B. subtilis*, our methodology provides an algorithmic approach for finding important sub-systems of organisms for which the GRN and MN has been obtained.

Finally we also attempt to find sub-modules of the largest SCC of *E. coli* which is large and complicated and cannot be assigned a simple functional role.

2 The GRN of *E. coli* and *B. subtilis*

We obtain the gene regulatory network interaction data of *E. coli* from RegulonDB [37], and that of *B. subtilis* from Freyre-Gonzalez et al [38] which is a curated database of regulatory interactions based on DBTBS, a database of transcriptional regulation in *B. subtilis* [39]. The GRN of *E. coli* contains 3277 nodes and 8740 edges. The GRN of *B. subtilis* consists of 1681 nodes and 3096 edges. Further details are given in Table 1. For reference purposes, we designate these GRNs (without consideration of feedback from metabolism) as graph $\mathcal{G}_{\mathcal{A}}$. The GRNs, without any knowledge of hierarchy, are pictured in Fig 1A. The entire data of graph $\mathcal{G}_{\mathcal{A}}$ (list of various types of nodes, links, hierarchical level, etc.) is given in S1 Table.

2.1 The hierarchical structure

The GRNs of bacteria are known to have a pyramidal hierarchical structure where the number of nodes in the hierarchical levels decreases as one moves to the higher levels [7, 9, 11, 13, 15, 27, 45]. However, with the growth in the size of the networks in recent years through the addition of new regulatory interactions, feedbacks have emerged. Feedbacks disturb the simple chain of command architecture and are critical because they give rise to properties like homeostasis, bistability etc. Our effort is directed towards the understanding of global organization and causal relationships between different regulatory elements of the GRN. Therefore we first remove the self loops so that we can focus on inter-nodal feedbacks that exist in the network. We capture the feedbacks in the network obtained after removing self loops by identifying its strongly connected components (SCCs). A SCC of a directed graph is a maximal subgraph, such that for any pair of nodes *i* and *j* in the subgraph, there exists a path in the sub-graph from *i* to *j* and from *j* to *i*. Together the SCCs capture all the feedbacks in the network. Furthermore, a SCC, owing to its property that any node belonging to it can affect any of its other nodes, can be thought of as forming a semi-autonomous group of nodes, or a module, of the full network. Understanding the functioning of these modules and placing them within the parent network can provide a way to understand the functional architecture of the parent network [14].

Varying procedures of defining the hierarchical levels for a GRN exist in the literature [7, 13, 15, 27]. Here, we use a combination of the procedure proposed by Jothi et al [13] along

Table 1. Networks overview. Overview of the various *E. coli* and *B. subtilis* networks used in the study.

		E. coli	*B. subtilis*
Gene Regulatory Network ($\mathcal{G}_{\mathcal{A}}$)			
	Total Genes	3277	1681
	Total regulator genes	248	154
	σ-factor genes	7	14
	Transcription factor genes	191	126
	Regulating ncRNA genes^{+}	50	14
	Regulated-only genes*	3029	1527
	Interactions	8740	3096
	Levels in Hierarchy	7	7
	Number of SCCs	7	4
	Size of largest SCC	56	6
Metabolic Network			
	Genes	904	1103
	Metabolites	761	1139
	Reactions	931	1437
GRN with allosteric feedbacks from Metabolic network ($\mathcal{G}_{\mathcal{B}}$)			
	Nodes	3343	1710
	Gene Nodes	3277	1681
	Metabolite Nodes	66	29
	Total edges	9279	3546
	Gene to gene edges	8740	3096
	Gene to metabolite edges	462	416
	Metabolite to gene edges	77	34
	Levels in Hierarchy	7	11
	Number of SCCs	20	9
	Size largest SCC	378	85

$^{+}$ ncRNA genes include both sRNA and tRNA genes.

* Genes whose product does not directly regulate any other gene (such nodes have no outgoing links in the GRN, but receive incoming links from regulator nodes.)

https://doi.org/10.1371/journal.pone.0203311.t001

with an intuitive bottom-up placement of hierarchical levels proposed by Yu et al [7], to construct the hierarchical levels (see Methods section 8.1 for details). Briefly, we perform a condensation of the GRNs of *E. coli* and *B. subtilis*, which produces corresponding directed acyclic graphs (DAGs). The condensation of a graph involves identification of SCCs of the graph, replacing each of the SCC by a single node (SCC node), and replacing the edges to/from the original nodes in the SCCs by new edges to/from the SCC node. The DAG, by construction, is devoid of cycles or feedback loops, and hence it is possible to unambiguously place the nodes of the network in a chain of command hierarchy. The hierarchical organization of the GRN of *E. coli* and *B. subtilis* is shown in Fig 1B.

The hierarchical organization of the GRNs obtained through the procedure above clearly elucidates the causal regulatory relationships between the transcriptional regulators as well as the regions of feedbacks (SCCs) in the GRNs. In the Fig 1B, all the regulatory links point downwards as a node in a higher level can regulate a node in a lower level but no node in the lower level can regulate a node in the higher level. This depicts the flow of information or regulatory influence in a GRN. We see that the hierarchy of transcriptional regulators in the GRNs have only a few localized islands of feedback.

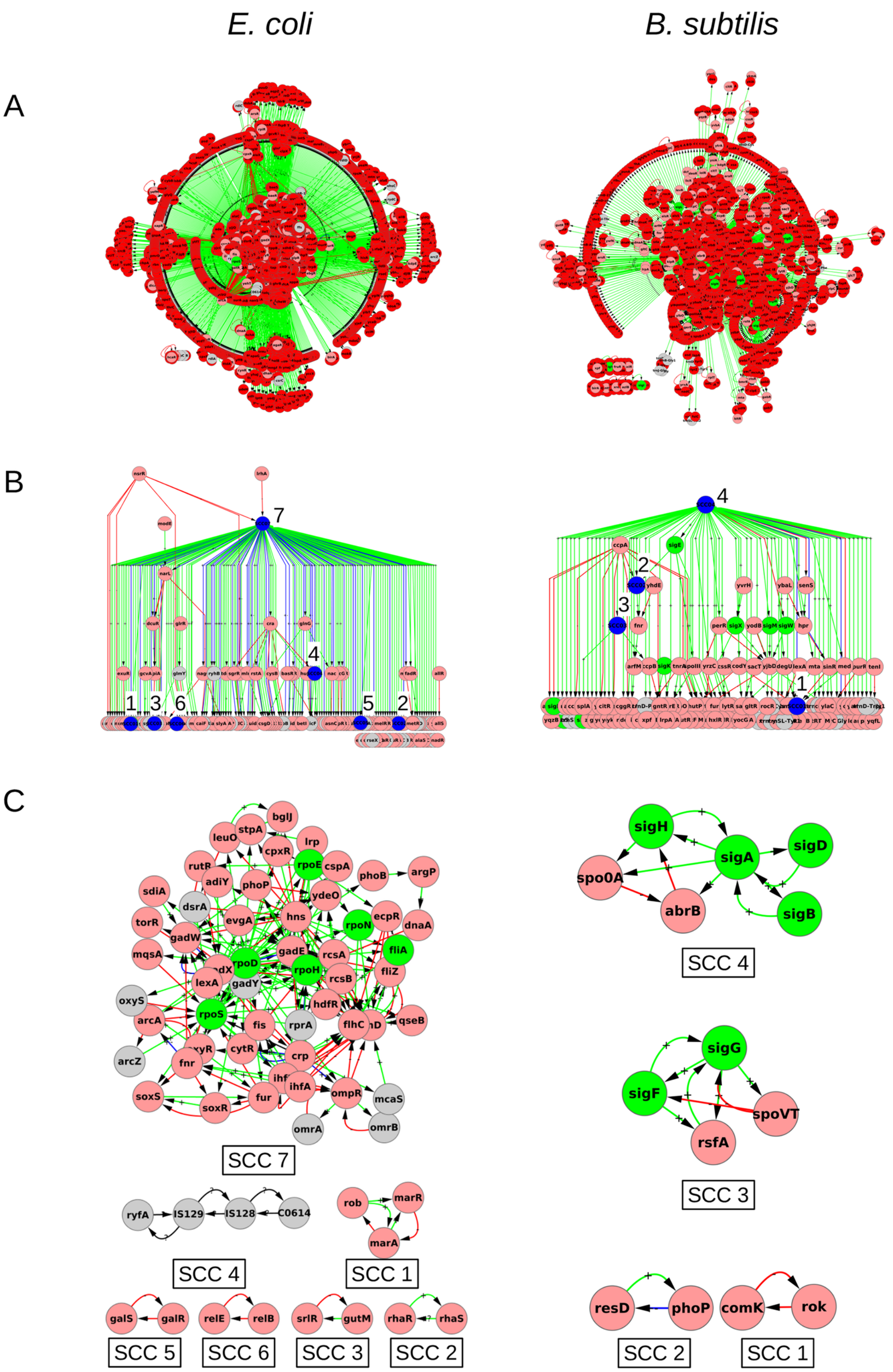


Fig 1. GRNs of *E. coli* and *B. subtilis*: Graph $\mathcal{G}_{\mathcal{A}}$. A. The whole GRNs of *E. coli* and *B. subtilis* pictured using one common layout. **B.** The revealed hierarchical structure (condensed graph) of the GRNs. Only the genetic regulators are shown. The regulated-only genes (nodes with no outgoing links) have not been shown for purposes of clarity; they would have been placed below the lowest level shown of the hierarchy, constituting a level-0. In other words, only level 1 and higher of the hierarchy are shown. Node colors—Blue: strongly connected components (SCCs), Pink:

transcription factor genes (TFs), Red: genes coding for enzymes, Green: σ-factor genes, grey: non-coding RNAs (ncRNAs). Edge colors represent the sign of regulation—Green (+): activating, Red (−): inhibiting, Blue (+− or −+): dual, Black (?): interaction mode remains uncharacterised. The sign (+, −, +−, ?) is also marked on the link. The SCCs have been numbered in the hierarchy, with their composition shown in C. **C.** The SCCs of the GRNs of *E. coli* and *B. subtilis*. Networks are pictured using Cytoscape [46]. The electronic versions of all figures in the paper can be zoomed in to read node names and the signs of links.

https://doi.org/10.1371/journal.pone.0203311.g001

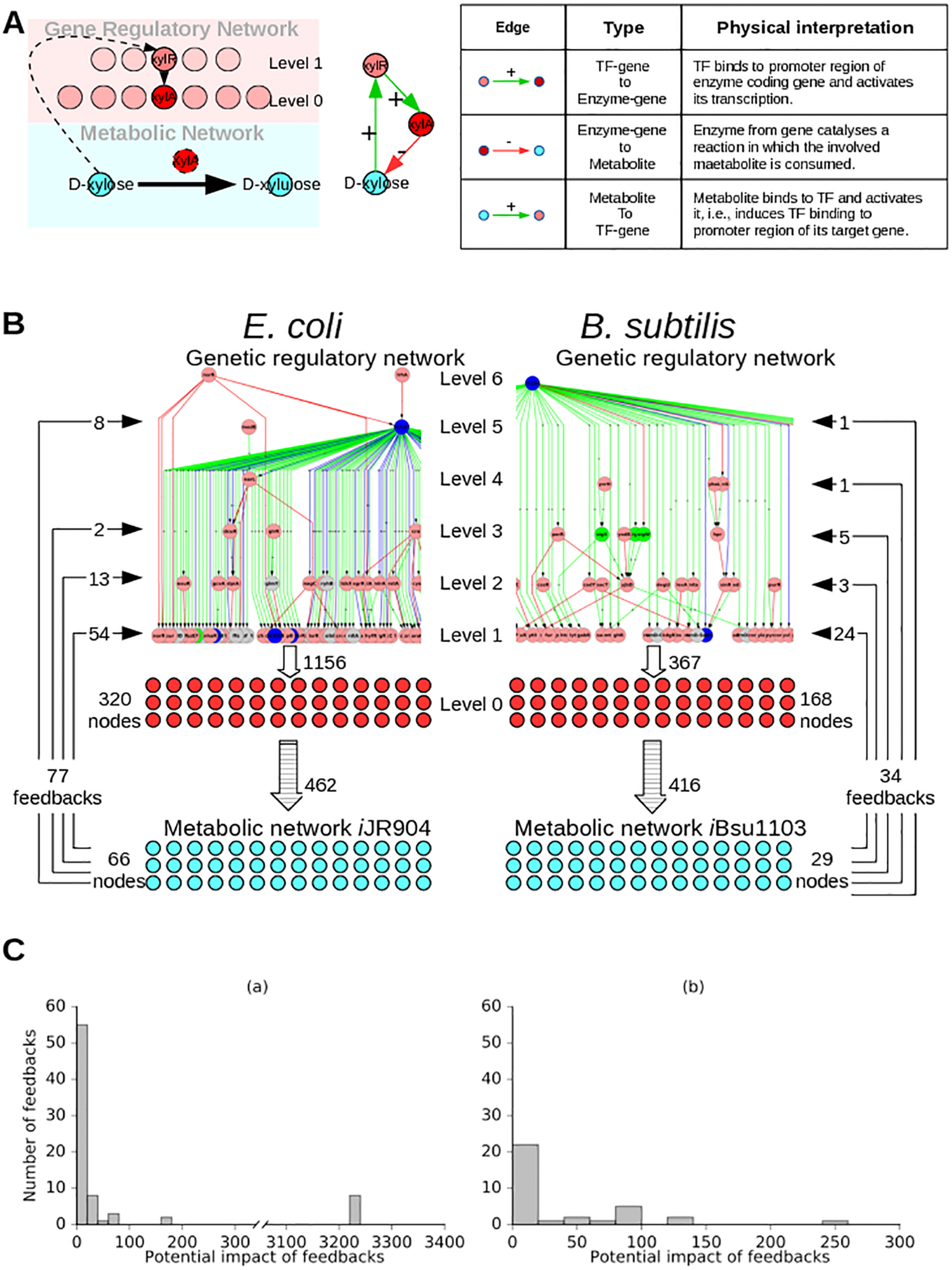
2.2 The inter-genic feedbacks

The GRN of *E. coli* has a total of 7 SCCs while the GRN of *B. subtilis* has a total of 4 SCCs (Fig 1C). Most SCCs in both organisms are small, having between 2 and 4 nodes. The largest SCC (LSCC) in *E. coli* (SCC 7 in Fig 1C) has 56 nodes, and is much larger than the second largest SCC which has only 4 nodes. In *B. subtilis*, however, the largest SCC has only 6 nodes and is not much larger than the second largest SCC. It is likely that this feature is a consequence of the fact that the present database of *B. subtilis* has far less coverage of the actual network compared to *E. coli*. It may be noted that only 1681 genes in *B. subtilis* are represented in the present network compared to 3277 in *E. coli* (see Table 1) while the total number of genes in both organisms is comparable and in the range 4000-4500. The average degree of the present GRN for *B. subtilis* (1.84) is also smaller than for *E. coli* (2.67) suggesting that many links may not have been documented. An earlier version of the *E. coli* network also did not have a giant SCC (the largest SCC in [14] had 11 nodes). Thus it is possible that with greater coverage of the *B. subtilis* network in the future a larger giant SCC may emerge in the *B. subtilis* network.

Despite these differences in size, the largest SCCs of both the organisms have two striking similarities. First, they are located at the top of the hierarchy. Second, both consist of global regulatory factors. The largest SCC in *E. coli* contains 6 (of 7) σ-factors and global TFs like *crp*, *fis*, *ihf*, *hns*. The largest SCC in *B. subtilis* contains the housekeeping sigma factor *sigA*. Quantitatively, the largest SCC of *E. coli* has over 80% of the top 10 global regulators (based upon outdegree of the genes), while *B. subtilis* has over 25%. It is clear from the respective set of SCCs that, for the present constructions of GRNs, the GRN of *E. coli* has more cycles than that of *B. subtilis* (on account of more SCCs and larger size of the LSCC). The above suggests that the GRNs of both bacteria have largely hierarchical tree like structures with mainly small and isolated feedback loops, and one larger feedback structure located near the top of the hierarchy that influences a large number of downstream genes. This picture is essentially the same as in [14] except that we find a much larger LSCC in *E. coli* at the top of the hierarchy. The LSCC in *E. coli* in the present study is 14 times larger than the second largest SCC. Some of the smaller SCCs in [14] now have coalesced into one and form part of the present LSCC.

3 Feedbacks from metabolic network into GRNs

The activity of some TFs is controlled by the binding of small molecules that are part of the cell's metabolism. These metabolites typically bind to the TF to alter its capacity to bind to target promoters. The production/consumption of such metabolites may be catalysed by enzymes whose genes are regulated by the same TF. Each binding reaction of a metabolite with a TF that alters the latter's regulatory activity thus creates a feedback from the MN to the GRN (see example in Fig 2A). We gather the information pertaining to the enzymatic catalysis of reaction from the metabolic models of the bacteria [40, 41], the TF metabolite interaction from RegulonDB [37, 47], Ecocyc [42] and Goelzer et al [43], and integrate this information with the GRN. For the purposes of reference, we label the GRN augmented with feedbacks from respective metabolic networks as graph $\mathcal{G}_B$ (details in Methods section 8.2; schematic in Fig 2B).



Edge	Type	Physical interpretation
○ —+→ ●	TF-gene to Enzyme-gene	TF binds to promoter region of enzyme coding gene and activates its transcription.
● —-→ ○	Enzyme-gene to Metabolite	Enzyme from gene catalyses a reaction in which the involved maetabolite is consumed.
○ —+→ ●	Metabolite To TF-gene	Metabolite binds to TF and activates it, i.e., induces TF binding to promoter region of its target gene.

Fig 2. Schematic of feedbacks into GRN from MN. A. Example of the feedback from the MN to the GRN. The TF *XylR* (coded for by the gene *xylR*), when activated, binds to and switches on the gene *xylA* (which codes for the enzyme *XylA*). *XylA* catalyzes the metabolic reaction in which metabolite D-xylose is converted into D-xylulose. D-xylose binds to *XylR* and activates it namely, causes it to bind to its target gene). The two figures in A are representation of the above statements. The table categorizes each type of link involved and gives its physical interpretation. This feedback has the dynamical consequence that excess of D-xylose in the cell is regulated by enhancing its

conversion into D-xylulose (through activation of *XylR* that up-regulates *XylA* which catalyzes the conversion). **B.** GRNs with feedbacks from metabolic network. **C.** Histograms (bin size 20) of potential impact of feedbacks from metabolic network into the respective GRNs of *E. coli*, (a), and *B. subtilis*, (b). Note that to avoid clutter we do not introduce separate nodes for a TF and the gene that codes for it. Thus the node representing a TF coding gene does double duty and represents both the genes and the corresponding TF. An arrow from a metabolite node (cyan) to a TF coding gene (pink) does not mean that the metabolite increases the expression of the gene. It means that the metabolite activates the TF coded for by the gene (namely, causes the TF to bind to its target genes and carry out its function—activating or inhibiting transcription of its target gene). However a green/red arrow (+ or −) from a TF coding gene (pink) to another TF coding gene (as in GRN of panel B) means that the former TF activates/inactivates the transcription of the latter gene.

https://doi.org/10.1371/journal.pone.0203311.g002

Fig 2B shows that most of the feedbacks from the metabolic network into GRN are at lower levels. This conclusion also holds for the number of feedbacks per transcription factor at each level. E.g., for *E. coli*, at levels 1 through 6, the number of feedbacks per TF at that level is 0.46, 0.50, 0.50, 0.00, 0.19, 0.00, respectively. Though the feedbacks at higher levels are fewer in number and in density, they impact a larger number of genes than the ones at the lower level. We explored the potential impact of feedbacks from metabolic network into GRN. One can define the potential impact of a feedback from a metabolite to a TF as just the number of genes downstream of that TF (at any distance). A histogram of the potential impact of feedbacks is shown in Fig 2C. As expected from Fig 2B and 2C shows that most of the allosteric feedbacks from metabolic networks into GRNs are employed to affect only a small set of genes, and a few feedbacks are also employed to bring about global changes. Interestingly, the feedbacks from metabolic networks into GRNs do not seem to have intermediate levels of potential impact (see Fig 2C). It would be interesting to await more data for *B. subtilis*, wherein the LSCC and perhaps feedback to it could emerge.

Feedbacks with large impact can possibly be related to effecting global changes, while those with low impact are likely to be effecting local changes. An example of a feedback having a global impact is from the metabolite cAMP (cyclic adenosine mono phosphate) which binds the TF *Crp* (coded by gene *crp*) in the core of the *E. coli* network (it is one of the eight mentioned in Fig 2B). It is well known that the absence of readily metabolizable carbon sources, such as glucose, results in the increased level of cAMP, which through its binding to *Crp* influences a large number of genes to change the state of the cell [48–50]. On the other hand the D-xylose feedback, described in Fig 2A, is at level 1. Its effect is more local. Instead of making a global impact, the feedback only affects a few genes like *xylA* involved in D-xylose metabolism with an effect to restore D-xylose balance. (A list of metabolites, the TFs they bind to and the level of the TF are given in S2 Table.) Level wise distribution of the feedbacks, Fig 2B and 2C, indicates that a majority of the allosteric feedbacks are utilized for local purposes.

4 Structure and organization of GRN with metabolic feedbacks

The incorporation of feedbacks into the GRN from the metabolic network brings substantial changes in the structure and organization of the GRN. For the purpose of reference, we label the GRN augmented with feedbacks from metabolism as graph $\mathcal{G}_{\mathcal{B}}$. We probe the structure and organization of graph $\mathcal{G}_{\mathcal{B}}$ in the same algorithmic way as done previously for $\mathcal{G}_{\mathcal{A}}$, i.e., via strongly connected components and hierarchical structure of the corresponding $\mathcal{G}_{\mathcal{B}}$ condensed graphs. We first describe the SCCs, and then the hierarchical structure. The graph $\mathcal{G}_{\mathcal{B}}$ is depicted in Figs 3 and 4.

4.1 Strongly connected components

Due to addition of feedbacks involving enzymes and metabolites, the largest SCCs were 6 fold and 14 fold bigger than SCCs of pure GRN (*E. coli* and *B. subtilis* respectively). For *E. coli*,

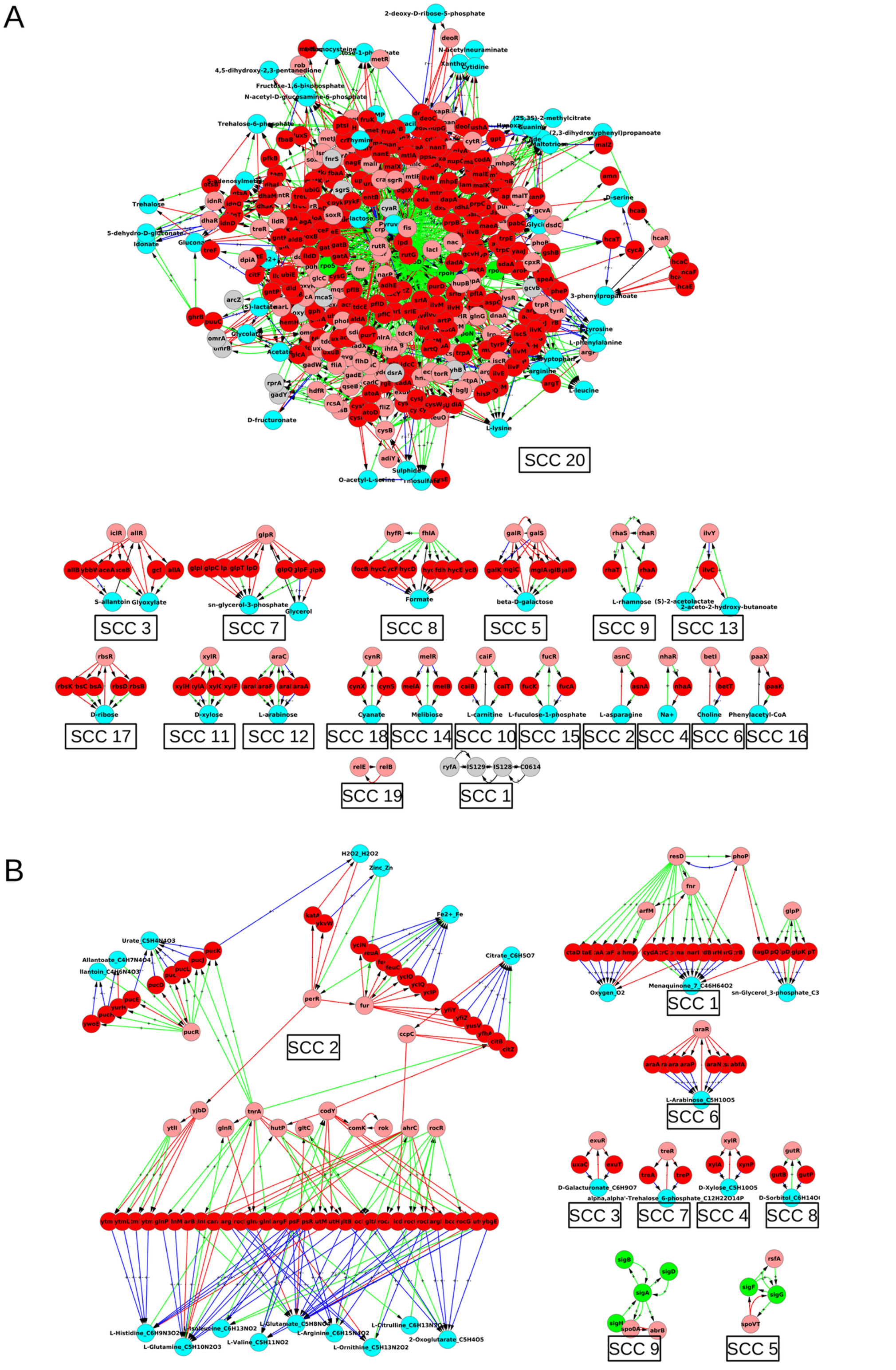


Fig 3. Strongly connected components of GRNs of *E. coli* (A) and *B. subtilis* (B) with feedbacks from respective metabolic networks: Graph $\mathcal{G}_B$. Colour code: Nodes—Red: genes coding for enzymes, Cyan: metabolites, rest same as in Fig 1; Edges—same as in Figs 1 and 2. In addition to the convention described in Fig 2A, we mention that a red arrow from a metabolite node to a TF coding gene node means that the metabolite binds to the TF and inactivates it, i.e., prevents the TF from binding to its target. Further a green arrow from an enzyme coding gene to a metabolite

means that the enzyme catalyzes a reaction in which the metabolite is produced. In addition to the convention described in Fig 2A, we mention that a red arrow from a metabolite node to a TF coding gene node means that the metabolite binds to the TF and inactivates it, i.e., prevents the TF from binding to its target. Further a green arrow from an enzyme coding gene to a metabolite means that the enzyme catalyzes a reaction in which the metabolite is produced.

https://doi.org/10.1371/journal.pone.0203311.g003

increase in number of feedbacks further complicates the already complex structure of the largest SCC. In both the organisms, the amino acids are predominantly present in the largest SCCs, whereas the sugars mostly occupy the small SCCs. This makes sense with the following logic: The sub-networks involving the sugars function at the input end of the metabolic network. When a sugar is present as food in the environment, the corresponding sub-network is active, while sub-networks related to other sugars are inactive. Thus the sub-networks relating to the sugar molecules form individual SCCs. Once the food is taken in, a core machinery which is active in every simulated environmental condition can produce various required molecules, e.g., the building block amino acids, for the cell. We will later (section 5) describe results which will shed more light on this observation. One stark difference between the largest SCCs of the two organisms relates to the presence of σ-factors in the LSCC of *E. coli* while the σ-factors are absent in the LSCC of *B. subtilis*. On the other hand there is also a similarity: amino acid and peroxide/iron modules are in the largest SCCs for both. For *B. subtilis* the new largest SCC can be loosely organized into 3 communities, Fig 3B, related to (a) one enriched in amino acids, (b) allantoin metabolism, and (c) with three sub-structures related to peroxide response, iron uptake and citrate metabolism. However, the LSCC (378 nodes) of *E. coli* is densely connected and its decomposition into communities is not that obvious. We discuss some aspects of this later (section 6). Further, the number of small SCCs has grown substantially both in size and in numbers (*E. coli*: 6 to 19, and *B. subtilis*: 4 to 8).

4.2 The hierarchical structure of graph $\mathcal{G}_{\mathcal{B}}$

With the inclusion of feedbacks into the GRNs of *E. coli* and *B. subtilis* from their associated metabolic networks, the enzyme coding genes and metabolites also get incorporated into the hierarchical structure, both as constituents of new SCCs and also singly. The levels in the hierarchy also get restructured. E.g., consider the gene *melR*. *melR* occurred in the level 1 of the hierarchy when the feedback from metabolism into the GRN were not considered and did not form a SCC, (visible upon zooming into Fig 1B for *E. coli*). In graph $\mathcal{G}_{\mathcal{B}}$ it not only forms a SCC (*SCC14*, Fig 3A) consisting of four nodes—1 gene coding for TF (*melR*), 2 genes coding for enzymes (*melA*, *melB*) and 1 metabolite (Melibiose)—but also shifts to level 2 (Fig 4A). While in $\mathcal{G}_{\mathcal{A}}$ it regulated only genes coding for enzymes, in graph $\mathcal{G}_{\mathcal{B}}$ it regulates SCC5 having 2 genes coding for TFs, 5 for enzymes, and 1 metabolite (see Figs 3A and 4A).

For *E. coli* graph $\mathcal{G}_{\mathcal{B}}$ the largest SCC (*SCC20*, Figs 4A and 3A) is built by addition of further nodes and links to the largest SCC (*SCC7*, Fig 1B and 1C) of *E. coli* graph $\mathcal{G}_{\mathcal{A}}$, and is similarly located at top but one level of the hierarchical organization. For *B. subtilis* graph $\mathcal{G}_{\mathcal{B}}$, the SCC at the top (*SCC9*, Figs 4B and 3A) is same as the one located at the top in case of graph $\mathcal{G}_{\mathcal{A}}$ (*SCC4*, Fig 1B), but is not the largest SCC. The largest SCC for the *B. subtilis* graph $\mathcal{G}_{\mathcal{B}}$ is *SCC2* which is located at level 4, and is heavily regulated (high in-degree), but regulates few (low out-degree), Figs 4B and 3B. This difference exists because amino acid related SCCs do not appear to interact with the SCC containing sigma factors. This in turn was because 4 out of 6 genes of the largest SCC of *B. subtilis* graph $\mathcal{G}_{\mathcal{A}}$ coded for σ-factors which generally are not modulated by the metabolites, and the other 2 genes coded for TFs received no feedback from metabolism. It would be interesting to see whether, as the database for *B. subtilis* expands, the amino

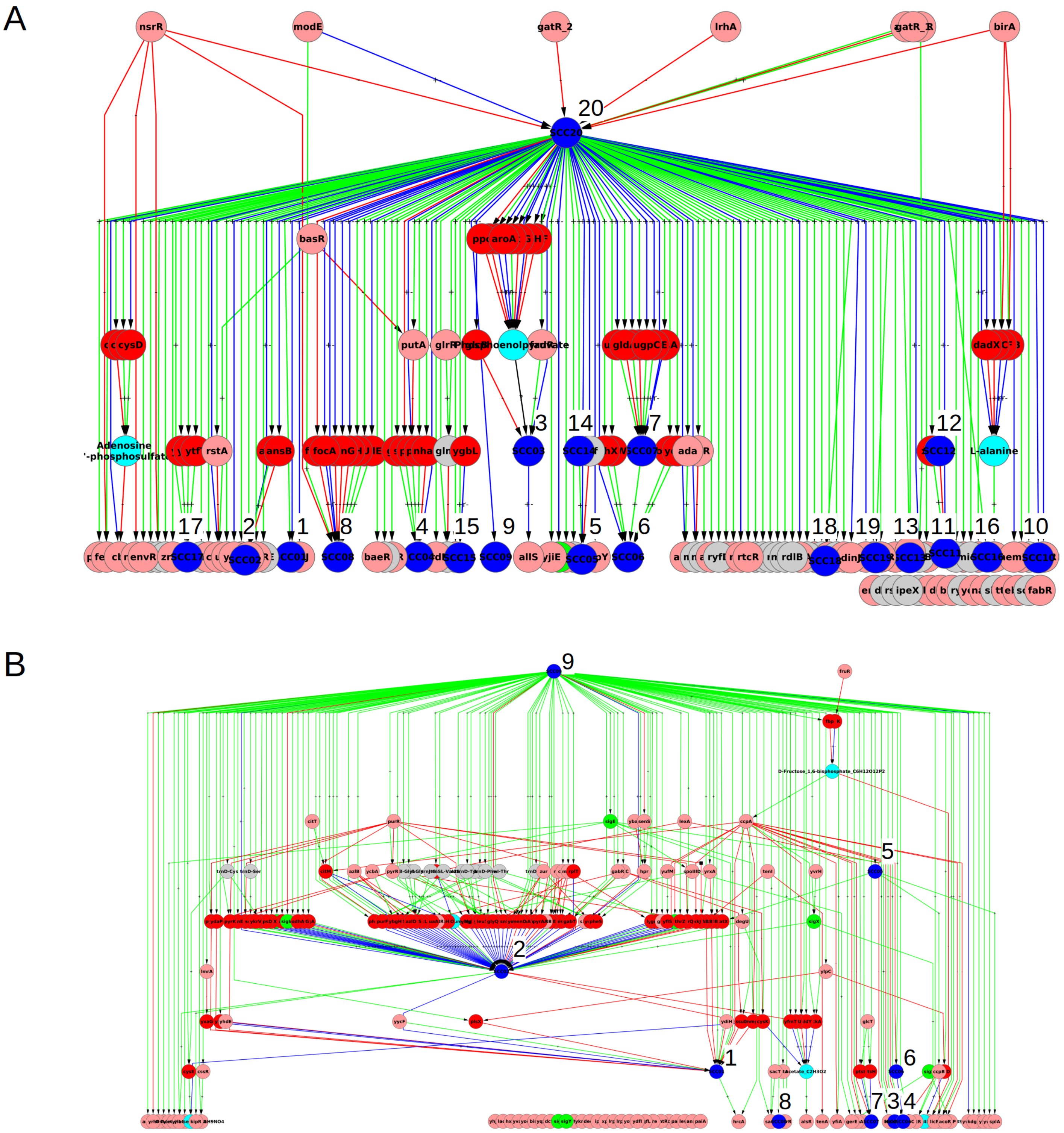


Fig 4. Hierarchical structure of GRNs of *E. coli* (A) and *B. subtilis* (B) with feedbacks from respective metabolic networks: Condensed version of graph $\mathcal{G}_{\mathcal{B}}$. The inclusion of feedbacks from metabolic network into respective GRNs reshuffles the hierarchy, without feedbacks ($\mathcal{G}_{\mathcal{A}}$), and also resolves the hierarchy further into more levels. The blue nodes, representing SCCs, are numbered and the detail of each numbered SCC is shown in Fig 3. Note that cyan nodes, coding for metabolites are different from blue nodes, representing SCCs. Nodes and edges follow the same colour code as Fig 3.

https://doi.org/10.1371/journal.pone.0203311.g004

acid metabolism cluster *SCC2* merges with the housekeeping σ-factor cluster *SCC9* (as is the case in *E. coli SCC20* in Fig 3), or whether they remain separate.

5 Modules in GRN augmented with functional feedbacks from MN

The augmentation of the GRN with feedbacks from the MN—graph $\mathcal{G}_B$—substantially increased the amount of feedbacks in it (compare Figs 3 and 1C). In order to understand the structure and role of these feedbacks more clearly we construct another graph, $\mathcal{G}_C$, which is a sparser version of $\mathcal{G}_B$. We use the knowledge that not all but only a part of the metabolic network is active under a given environmental condition (EC) to simplify the augmented GRN by considering only essential feedbacks under a set of defined ECs. This approach allows us to study the functioning of the SCCs with regard to the given environment. In order to simulate various ECs, we use the computational technique of flux balance analysis (FBA) [44, 51] (see Methods 8.3 and 8.4). The ECs we consider are minimal media characterized by a single organic source of carbon and a few other essential metabolites. For the present metabolic models, we obtain a list of 158 ECs for *E. coli* and 118 for *B. subtilis*, in which the organism has a positive growth rate (see S3 Table for details). We determine essential reactions under each given EC, and consider the feedbacks into the GRN from the metabolites taking part in these reactions only. It is to be noted that the set of essential reactions is a subset of the set of functional reactions in the flux vector returned by FBA. However, the flux vector returned by FBA is not unique [44], i.e., different sets of active reactions exist that give the same growth rate for a given medium. On the other hand the set of essential reactions is unique for any medium. In this work we consider essential reactions in order to get unambiguous and easily reproducible results, albeit at the cost of missing out some functional reactions. Next, we augment the GRN with these feedbacks from the MN. We refer to this augmented version of the GRN as graph $\mathcal{G}_C$ (details in Methods section 8.3; detailed data of $\mathcal{G}_C$ in S4 Table; summary in Table 2).

A comparison of Table 2 with Table 1 shows that the number of metabolite nodes eliminated from *E. coli* and *B. subtilis* networks is 15 and 5 respectively, while the number of links eliminated is 250 and 208 respectively. The procedure (of going from $\mathcal{G}_B$ to $\mathcal{G}_C$) eliminates a substantial number of links. Thus while $\mathcal{G}_B$ and $\mathcal{G}_C$ both incorporate feedbacks from the metabolic network into the GRN, we can say that $\mathcal{G}_C$ includes only "functionally relevant" feedbacks, since it includes only those reactions and metabolites of $\mathcal{G}_B$ that are essential for the growth of the organism in some medium. Next, we analyze the logic of each SCC of $\mathcal{G}_C$ in detail in terms of its potential influence on the dynamics of the system by considering the signs of the links, the embedding of the SCC in the larger network, and the conditions in which the

Table 2. GRN augmented with functional feedbacks from metabolic network ($\mathcal{G}_C$). Details of composition of GRN augmented with feedbacks from metabolic network—graph $\mathcal{G}_C$—for *E. coli* and *B. subtilis*.

Feature	*E. coli*	*B. subtilis*
Nodes	3328	1705
Genes	3277	1681
Metabolites	51	24
Total edges	9029	3338
Gene to gene edges	8740	3096
Gene to metabolite edges	229	217
Metabolite to gene edges	60	25
Levels in Hierarchy	12	17
Number of SCCs	28	14
Size of largest SCC	97	13

https://doi.org/10.1371/journal.pone.0203311.t002

metabolic reaction link in the SCC is found to be active. This analysis shows that almost every SCC has a specific functional role in the organism and can be associated with a functional module of the system.

5.1 Strongly connected components and hierarchical structure of graph $\mathcal{G}_C$

The SCCs and condensed graph of $\mathcal{G}_C$ for *E. coli* are given in Figs 5, 6 and S1 Fig, and those of *B. subtilis* in Fig 7 and S2 Fig. The data of graph G_C for both organisms, is given in S4 Table, including nodes, links, distribution of nodes in hierarchical levels, details of each SCC, etc. There are more SCCs in graph $\mathcal{G}_C$ than in graph $\mathcal{G}_B$ (compare the numbers in Tables 1 and 2). The size of largest SCC of graph $\mathcal{G}_C$ for *E. coli* is considerably reduced (SCC 28 in Fig 5 has 97 nodes as compared to SCC 20 in Fig 3 having 378 nodes). Many sub-networks of the largest SCC of $\mathcal{G}_B$ now break out into individual smaller SCCs of $\mathcal{G}_C$ (SCCs 3, 5, 6, 7, 8, 9, 11, 14, 15, 16, 17, 20, 22, 23, and 27 in Fig 6) suggesting that the latter modules function independently in the conditions considered here. Some of the SCCs of $\mathcal{G}_B$ (SCC 2, 4, 6, 8, 10, 16, and 18) are absent in $\mathcal{G}_C$ because the corresponding metabolite is not a minimal medium food source in our considered ECs. For *B. subtilis* a similar phenomenon happens and the largest SCC of $\mathcal{G}_C$ is of size only 13. The LSCC of graph $\mathcal{G}_B$ (size 85; SCC 2 in Fig 3B) breaks up into smaller SCCs (SCC 1, 2, 3, 4 and 11 in Fig 7) of its constituent communities of allantoin, amino acids and citrate modules. The number of levels in the condensed graph of $\mathcal{G}_C$ for *E. coli* and *B. subtilis* are 12 and 17 respectively, S1 and S2 Figs.

It has been observed that real biochemical networks, e.g., protein-reaction networks [10] including genetic regulatory networks without feedback from metabolism [45], have a larger number of strong components than their randomized counterparts, thereby exhibiting a higher level of modularity. We find that the genetic regulatory graphs $\mathcal{G}_B$, $\mathcal{G}_C$ which contain feedbacks from metabolism into the GRN, when compared to their appropriately randomized versions, also have the same feature. Details are given in S7 Table.

5.2 Architecture and functionality of modules

The question arises as to what is the significance of the non-trivial SCCs. The dynamics of an SCC by definition depends upon the state of the nodes above it in the GRN hierarchy and on its own non-trivial internal structure. To study their functional significance we examine the internal structure of the SCCs obtained above for *E. coli* and *B. subtilis* in detail. We show below that almost every SCC obtained above is associated with an identifiable and definite functionality, or that it performs a specific task or set of tasks in the organism. It may be noted that a priori, it is not obvious that a non-trivial SCC should have a specific biological role. The fact that we find most SCCs to have a specific role suggests that this is a useful construct. The twin properties of being dynamically relatively autonomous and having an identifiable biological functionality justify our referring to these SCCs as associated with 'modules' of the GRN. Many algorithmic methods exist in the literature of identifying modules, e.g., those that cluster expression data and those that identify 'communities' in networks. Our algorithmic method is different, and it does not identify all modules in the system, but those that it does have a fairly tight functional significance, as will be seen below. Thus the method would be particularly useful in identifying modules for those organisms for which the genetic and metabolic databases will become available, but whose biology has not been as extensively studied as *E. coli* and *B. subtilis*.

The SCCs of Figs 6 and 7 have been displayed along with their properties in S5 and S6 Tables. We have focused on SCCs that have at least one metabolite. The total number of such

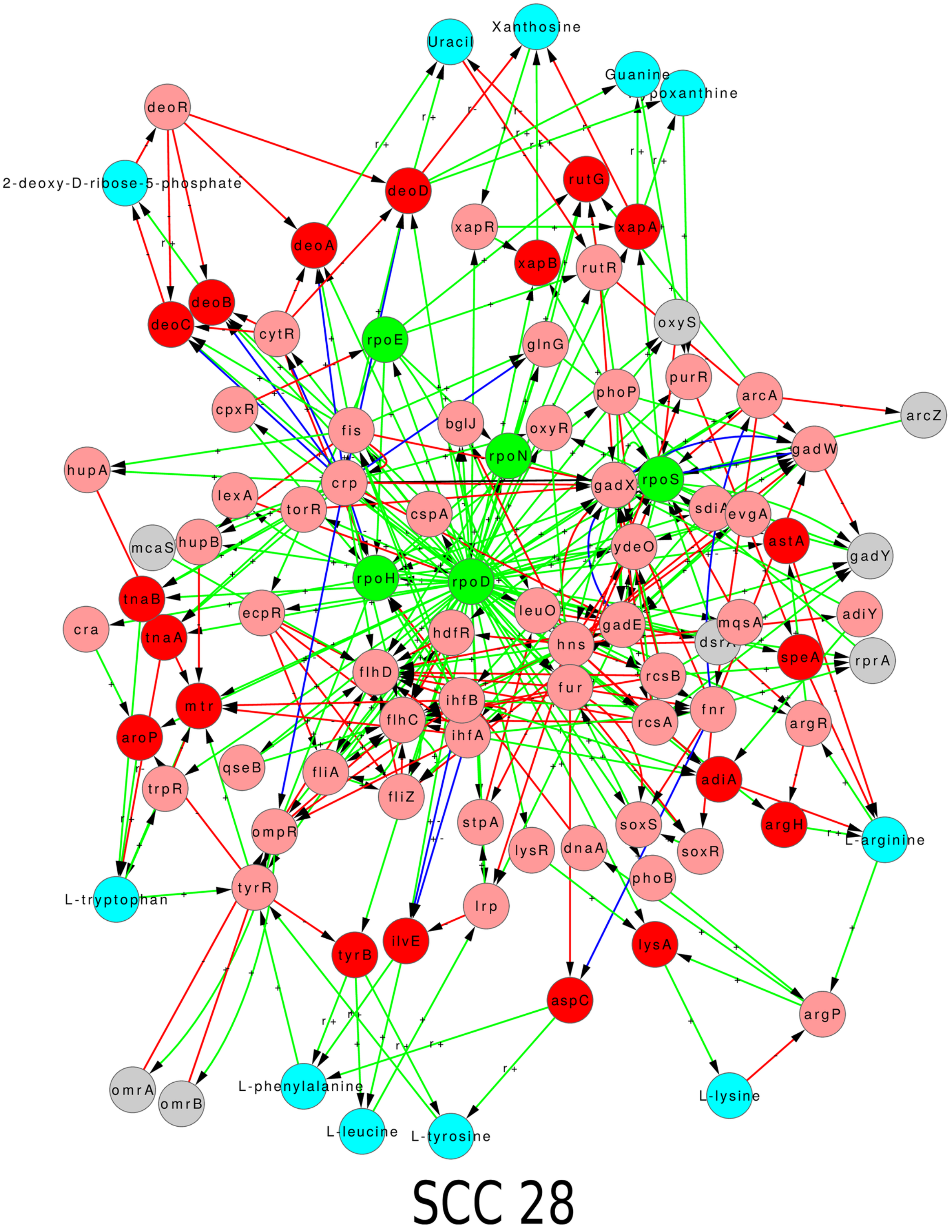


Fig 5. Largest strongly connected component of *E. coli* graph $\mathcal{G}_{\mathcal{C}}$ (GRN augmented with functionally relevant feedbacks from metabolic network). Nodes and edges follow the same colour code as in Fig 3.

https://doi.org/10.1371/journal.pone.0203311.g005

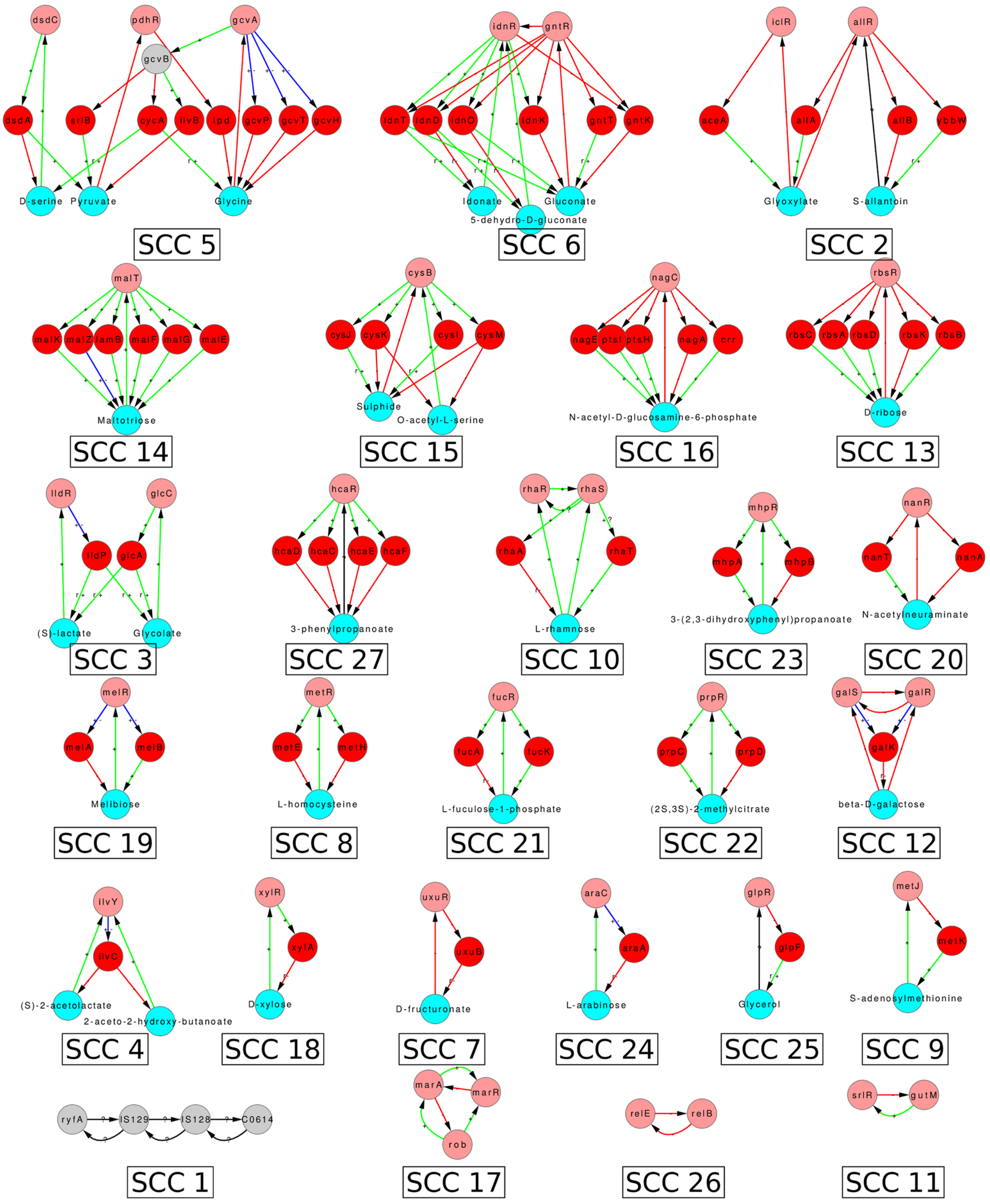


Fig 6. Smaller strongly connected components of *E. coli* graph $\mathcal{G}_C$. Nodes and edges follow the same colour code as in Fig 3.

https://doi.org/10.1371/journal.pone.0203311.g006

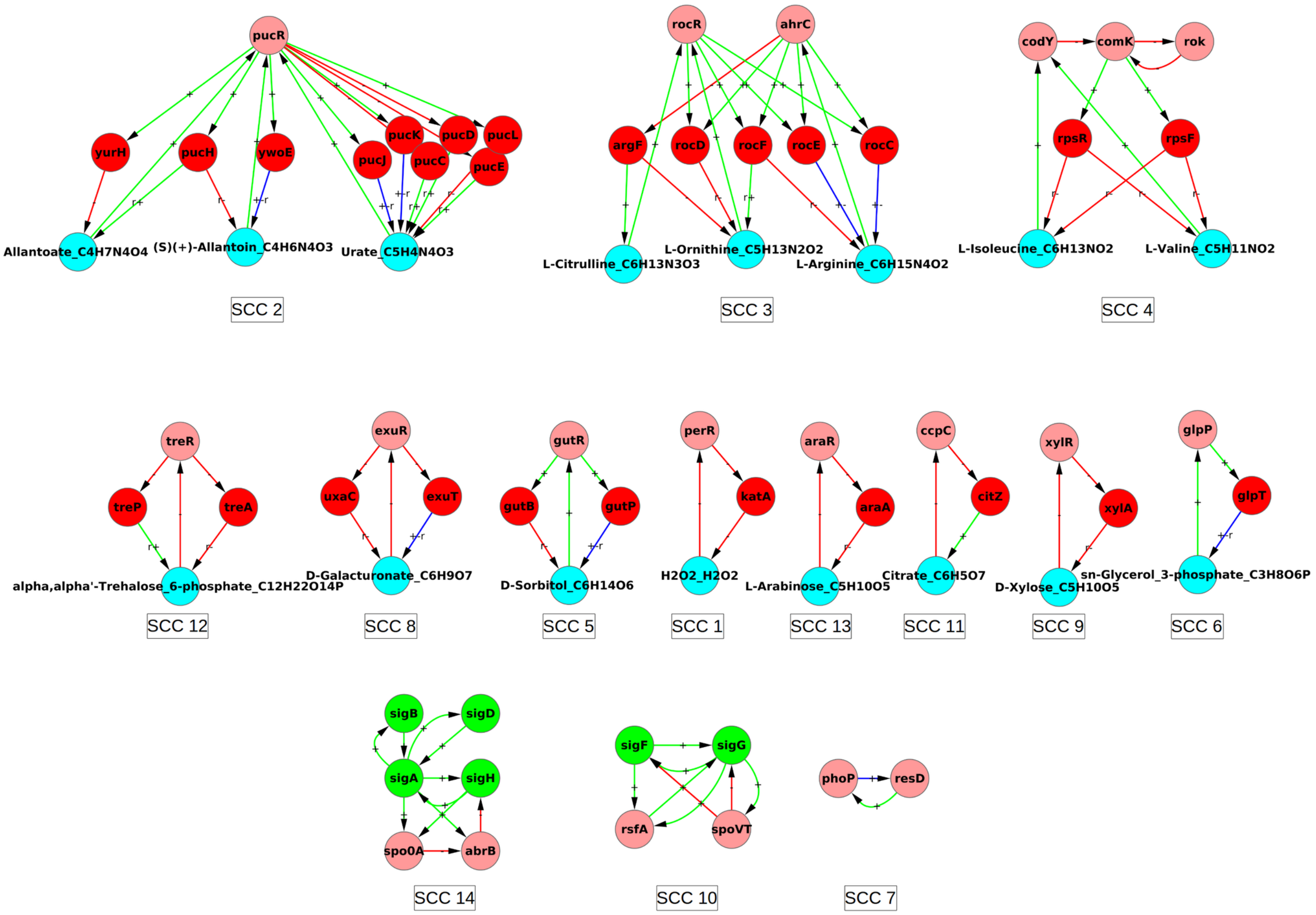


Fig 7. Strongly connected components of GRN of *B. subtilis* graph $\mathcal{G}_c$. The resulting SCCs of *B. subtilis* graph $\mathcal{G}_c$ are enumerated. Nodes and edges follow the same colour code as in Fig 4.

https://doi.org/10.1371/journal.pone.0203311.g007

SCCs is 23 in *E. coli* and 11 in *B. subtilis*. For *E. coli* the largest SCC, SCC 28 in Fig 5, is complicated, has multiple functional tasks, and is discussed separately in a later subsection.

It is convenient to organize the functional study of the metabolite containing SCCs, hereafter referred to interchangeably as modules, in ascending order of size (number of nodes), as the functionality of larger modules can often be understood in terms of smaller structures.

5.2.1 Size 3 modules: NFL or PFL. There are five size 3 modules in *E. coli* and five in *B. subtilis*. Each size 3 module has one metabolite, one TF gene and one enzyme gene. The modules of size 3 can be classified into two categories: negative feedback loop and positive feedback loop (NFL, PFL). While the composite logic in these modules remain either a NFL or PFL, their internal composition in terms of position and sign of the interaction may vary. Fig 8 tabulates examples of this structural variation that we find and details about each example displayed that leads to a statement about the possible functionality of the module. All the modules are shown in S5 and S6 Tables. As an example, we discuss the module in the first row of Fig 8. This module is found to be active only when the food source is Aldehydo-D-glucoronate. In that situation enzyme *UxaC* converts this molecule to D-fructuronate. One can guess that the function of the module, a NFL with three inhibitory links, is to metabolize D-fructuronate when it is in excess into a further downstream product. The internal structure of the module

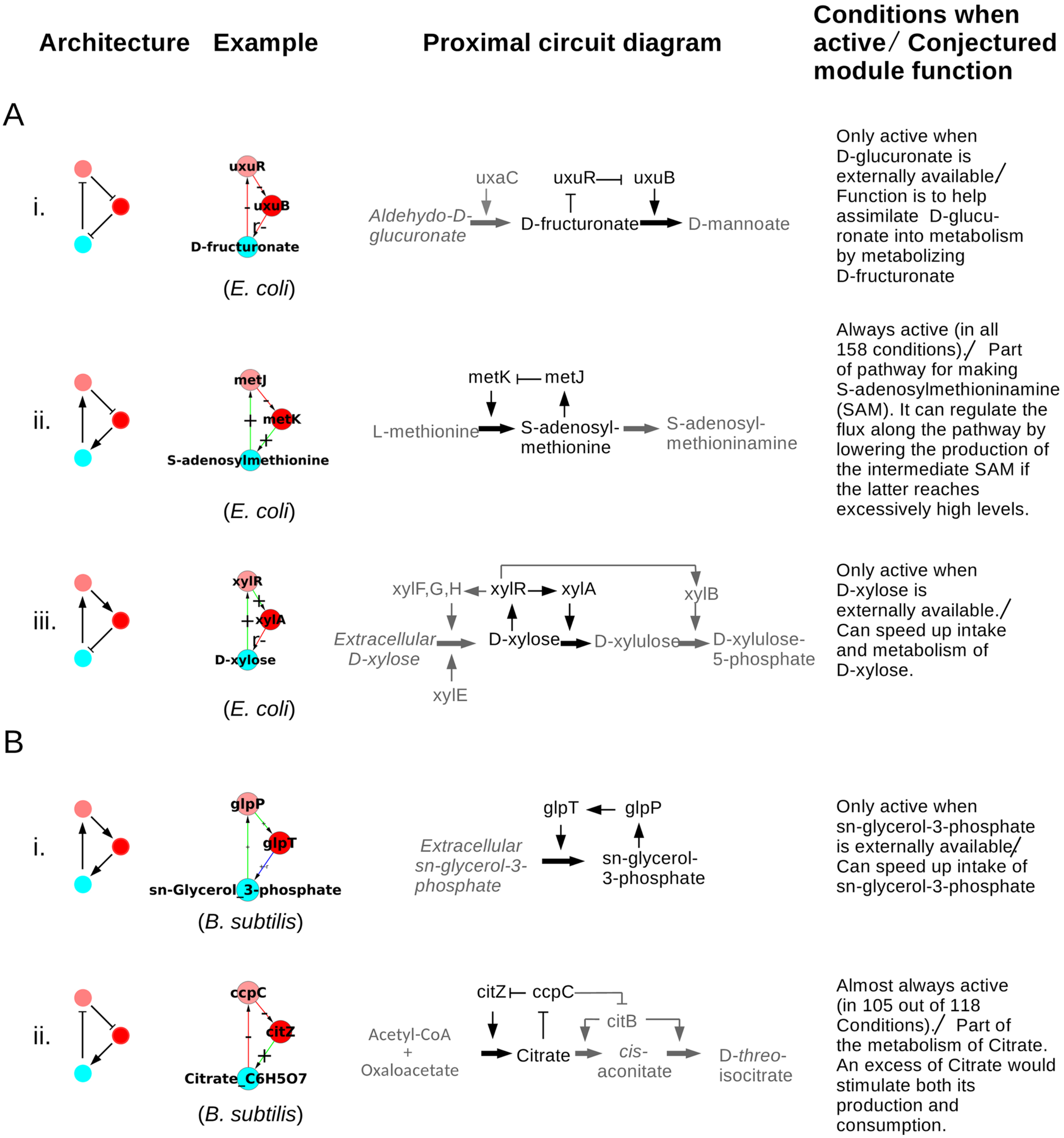

Architecture
Example
Proximal circuit diagram
Conditions when active/ Conjectured module function
A
i.
uxuR
uxuB
D-fructuronate
(E. coli)
uxaC
uxuR
uxuB
Aldehydo-D-glucuronate
D-fructuronate
D-mannoate
Only active when D-glucuronate is externally available/ Function is to help assimilate D-glucuronate into metabolism by metabolizing D-fructuronate
ii.
metJ
metK
S-adenosylmethionine
(E. coli)
metK
metJ
L-methionine
S-adenosyl-methionine
S-adenosyl-methioninamine
Always active (in all 158 conditions)./ Part of pathway for making S-adenosylmethioninamine (SAM). It can regulate the flux along the pathway by lowering the production of the intermediate SAM if the latter reaches excessively high levels.
iii.
xylR
xylA
D-xylose
(E. coli)
xylF,G,H
xylR
xylA
xylB
Extracellular D-xylose
D-xylose
D-xylulose
D-xylulose-5-phosphate
xylE
Only active when D-xylose is externally available./ Can speed up intake and metabolism of D-xylose.
B
i.
glpP
glpT
sn-Glycerol_3-phosphate
(B. subtilis)
glpT
glpP
Extracellular sn-glycerol-3-phosphate
sn-glycerol-3-phosphate
Only active when sn-glycerol-3-phosphate is externally available/ Can speed up intake of sn-glycerol-3-phosphate
ii.
ccpC
citZ
Citrate_C6H5O7
(B. subtilis)
citZ
ccpC
citB
Acetyl-CoA + Oxaloacetate
Citrate
cis-aconitate
D-threo-isocitrate
Almost always active (in 105 out of 118 Conditions)./ Part of the metabolism of Citrate. An excess of Citrate would stimulate both its production and consumption.

Fig 8. Architecture and function of 3-node modules. A. The negative feedback loop (NFL). **B.** The positive feedback loop (PFL). Beside the architecture an example of the same is shown, as well its proximal circuit diagram showing the elementary pathway logic of the involved genes and metabolites. In the latter diagram the genes and metabolites of the example module are shown in black, while the preceding and following elements of the pathway have been greyed, for distinction and clarity. Thick arrows represent metabolic reactions or conversion of one metabolite into another. An arrow from an enzyme coding gene to a thick arrow means that the enzyme catalyzes the reaction. Other arrows follow the same convention as in Figs 2 and 3. A flat tipped link from a metabolite to a TF means that when the metabolite binds to the TF it inactivates it, i.e., prevents it from binding to its target gene. We will say that an SCC is 'active' in an EC if a reaction (enzyme gene to metabolite link) belonging to the SCC is essential in that EC. The last column gives the environmental conditions in which the SCC is active, and the conjectured functional role of the module.

https://doi.org/10.1371/journal.pone.0203311.g008

shows how this can be achieved. D-fructuronate binds to the TF *UxuR* and inactivates it. *UxuR* was repressing the expression of gene *uxuB* and hence preventing the production of enzyme *UxuB*. When *UxuR* is inactivated by D-fructuronate, the enzyme *UxuB* is produced and this in turn catalyzes a reaction that consumes D-fructuronate. This functionality is useful when D-fructuronate is in excess, consistent with the fact that we only see it when aldehydo-D-glucoronate is the food source. It helps the assimilation of this food source into the metabolism. Further it prevents gene expression (which is costly [52]) of *uxuB* when not required, i.e., when the food source is not available.

Almost every topology we have found as a module has a clearly identifiable functionality or purpose which is evident from Fig 8. Broadly, two functions are served. (1) The modules of D-glucoronate, D-xylose and sn-Glycerol-3-phosphate—Fig 8A(i), 8A(iii) and 8B(i) respectively—are only active when the respective metabolite is itself the food source as the growth media, else they are inactive. These modules are present at the input end of the metabolic network. Their specific activity indicates that they function for the uptake or metabolism of respective metabolite food molecule in the cell. (2) The modules which are active in all or almost all of the growth conditions and are located deep in some pathway of the metabolite, e.g. the modules of S-adenosylmethionine and citrate, Fig 8A(ii) and 8B(ii). They can have different functions depending upon the structure. E.g., the S-adenosylmethionine module, Fig 8A (ii), seems to be designed to regulate the production (over/under- production) of S-adenosylmethionine, whereas the citrate module, Fig 8B(ii), seems to be designed to speed up the production of citrate. We note that out of the ten size-3 modules, four belong to category A(i) and one each to A(ii), A(iii), B(i) and B(ii). Of the remaining two, in one case the link from the TF to the enzyme is stated to have a dual sign (+ or -) in the database, and for the other there is some ambiguity about the existence of the link from the metabolite to the TF. Details are in the S5 and S6 Tables.

It is important to point out that the SCCs we find often uncover a larger module of which they are a part. The larger module can be constructed once the SCC is identified by our method. An example is the D-xylose module in Fig 8A(iii). The SCC found by our method only had the 3 nodes: the metabolite D-xylose, the TF *XylR* and the enzyme *XylA*. However the proximal circuit shows that *XylR* also activates enzymes *XylF,G,H* which catalyze the intake of extracellular D-xylose and *XylB* which converts D-xylulose into D-xylulose-5-phosphate. The *XylF,G,H* do not appear as a module in our procedure because its intake reaction is outside the scope our procedure as it is not an essential reaction because another enzyme *XylE* provides a second pathway for the intake of extracellular D-xylose. Collectively this whole system should be considered the “D-xylose module” instead of the 3 node SCC found by our method. Thus our method sometimes detects not the whole module but what might be called as ‘kernel’ of the module from which the rest of the module can often be constructed by examining the proximal circuit. We could have discovered that *XylB* is also part of the SCC if we had been more liberal in including metabolic reactions that produce or consume not just metabolites that bind to TFs but also their nearest neighbour metabolites. It is a task for the future to extend our methodology to include these effects, and to see the usefulness of the results so obtained.

The D-xylose module also illustrates another interesting feature: the overall function is conserved across organisms in spite of the variation in the internal structure of the module. D-xylose forms a size-3 SCC in both *E. coli* and *B. subtilis*, Fig 9. In both cases the module is a negative feedback loop (NFL). Also, in both cases, the module is active only when D-xylose is the food source and can speed up its intake and metabolism. However, the D-xylose module in *E. coli* has a different internal structure than that in *B. subtilis*. In *E. coli* the structure of the loop is *XylR* → *XylA* ⊣ D-xylose → *XylR* (an overall NFL). Whereas in *B. subtilis* the structure is *XylR* ⊣ *XylA* ⊣ D-xylose ⊣ *XylR* (also an overall NFL). In *E. coli* the D-xylose activates the TF

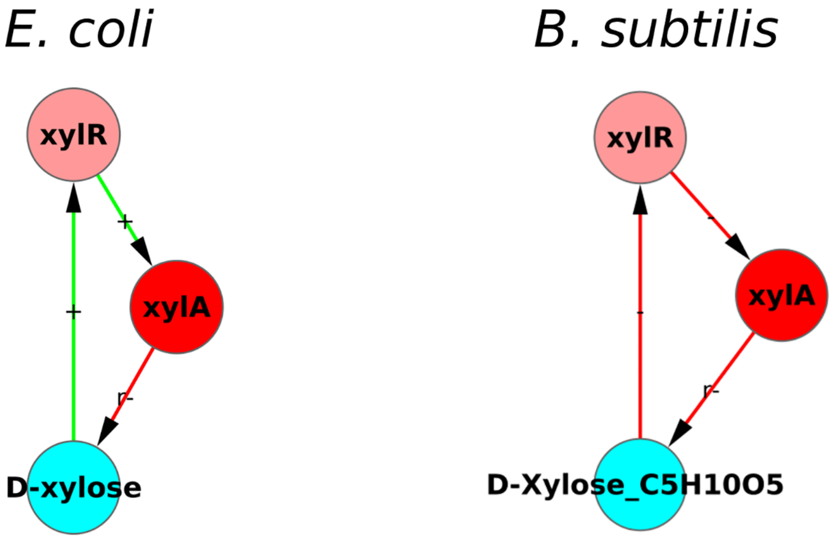


Fig 9. The size-3 D-xylose module (NFL) with same function but different internal structure. The D-xylose module in both *E. coli* and *B. subtilis* is an over-all negative feedback loop serving the same function of intake and metabolism of D-xylose, but has different internal structures in the two organisms. This is an example of the conservation of function of a module despite variations in its internal structure.

https://doi.org/10.1371/journal.pone.0203311.g009

XylR whose action is activation of enzyme *XylA* thus leading to the metabolism of D-xylose into Xylulose. Whereas in *B. subtilis* D-xylose is also metabolized into Xylulose by the enzyme *XylA* but through a different regulation: D-xylose inactivates the TF *XylR* whose action is repression of the enzyme *XylA*, with the overall effect the enzyme *XylA* again metabolizing D-xylose to Xylulose. Different ways of implementing the same logic can have similar effects in a steady state but could differ in transient effects.

5.2.2 Size 4 modules: Consumer motifs. The modules of size 4 can be broadly divided into two classes: (a) a 4-node motif associated with uptake and utilization of small molecules, (b) two essentially 3-node motifs sharing nodes and therefore forming a 4-node SCC. In the former case the full 4-node structure is essential to understand its functionality. In the latter case the individual 3-node sub-structures have a functionality in their own right.

The first class (Fig 10A) is a structure suitable for the uptake and metabolism of food molecule, studied in [53] and referred to as a 'consumer motif'. The consumer motif can be thought of as a combination of two 3-node motifs—PFL followed by NFL—where both the loops work in tandem for the uptake and metabolism of the food molecule. In a consumer motif the first loop serves to transport the metabolite into the cell from the outside, i.e., increase its cellular concentration, while the second loop metabolizes it to other compounds thereby decreasing its cellular concentration. We found that the majority of the 4-node motifs (6 out of 8 in *E. coli* and all 3 in *B. subtilis*, hence 9 out of a total of 11 in both organisms) are of consumer type. Example of both the variants of the consumer motif, one where the metabolite activates the TF (the sorbitol module) and the other where the metabolite inactivates the TF (the N-acetylneuraminate module), are shown in Fig 10A. We note that in most cases the consumer motif is located at the uptake end of a metabolite's pathway—6 out of 9 are involved in uptake and further processing of a metabolite. Of the remaining three, two are not involved in uptake but are close enough to the uptake reaction for their role to be reasonably clear (the dihydroxyphenol propanoate module and fuculose module). All these 8 modules are active with only one or a few very specific food sources. One (the methylcitrate module) is not located immediately

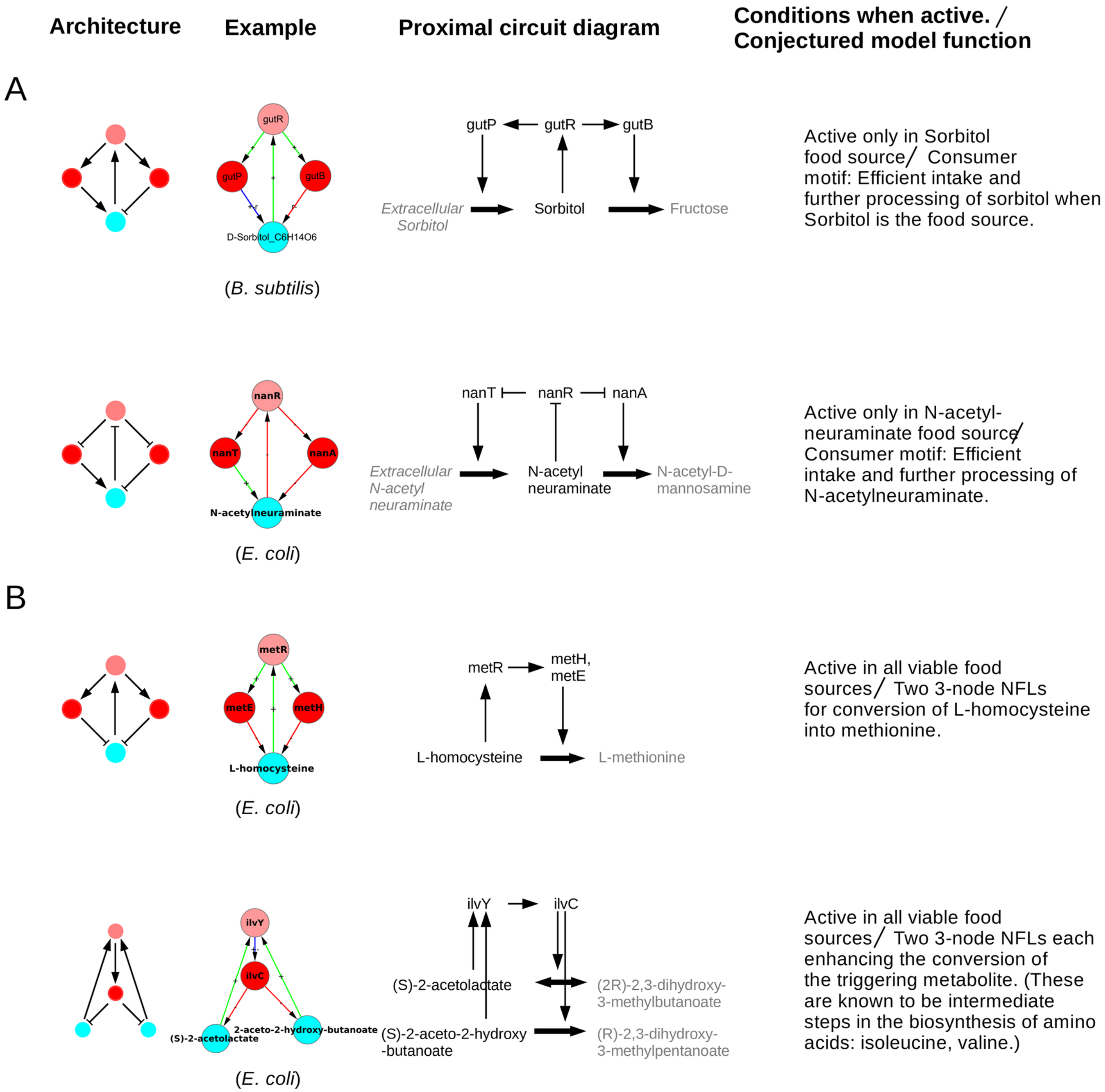

Architecture
Example
Proximal circuit diagram
Conditions when active. / Conjectured model function
A
gutR
gutP
gutB
D-Sorbitol_C6H14O6
(B. subtilis)
gutP
gutR
gutB
Extracellular Sorbitol
Sorbitol
Fructose
Active only in Sorbitol food source/ Consumer motif: Efficient intake and further processing of sorbitol when Sorbitol is the food source.
nanR
nanT
nanA
N-acetylneuraminate
(E. coli)
nanT
nanR
nanA
Extracellular N-acetyl neuraminate
N-acetyl neuraminate
N-acetyl-D-mannosamine
Active only in N-acetyl-neuraminate food source/ Consumer motif: Efficient intake and further processing of N-acetylneuraminate.
B
metR
metE
metH
L-homocysteine
(E. coli)
metR
metH, metE
L-homocysteine
L-methionine
Active in all viable food sources/ Two 3-node NFLs for conversion of L-homocysteine into methionine.
ilvY
ilvC
(S)-2-acetolactate
2-aceto-2-hydroxy-butanoate
(E. coli)
ilvY
ilvC
(S)-2-acetolactate
(S)-2-aceto-2-hydroxy -butanoate
(2R)-2,3-dihydroxy-3-methylbutanoate
(R)-2,3-dihydroxy-3-methylpentanoate
Active in all viable food sources/ Two 3-node NFLs each enhancing the conversion of the triggering metabolite. (These are known to be intermediate steps in the biosynthesis of amino acids: isoleucine, valine.)

Fig 10. Architecture and function of 4-node modules. A. The consumer motif along with its two identified variations. **B.** The 4-node modules which can be reduced to essentially two 3-node NFLs with shared nodes. Beside the architecture an example of the same is shown, as well as the elementary pathway logic of the genes and metabolites involved in the module, and functionality of the structure. Nomenclature is as in Fig 8.

https://doi.org/10.1371/journal.pone.0203311.g010

following external metabolite. Again its function is clear from the structure, but the reason for its deep location in the network is not clear (see S5 and S6 Tables for details).

In the second class (Fig 10B), the functionally non-reducible modules involved are the 3-node motifs. The only two cases in this class, the homocysteine module and the acetolactate-

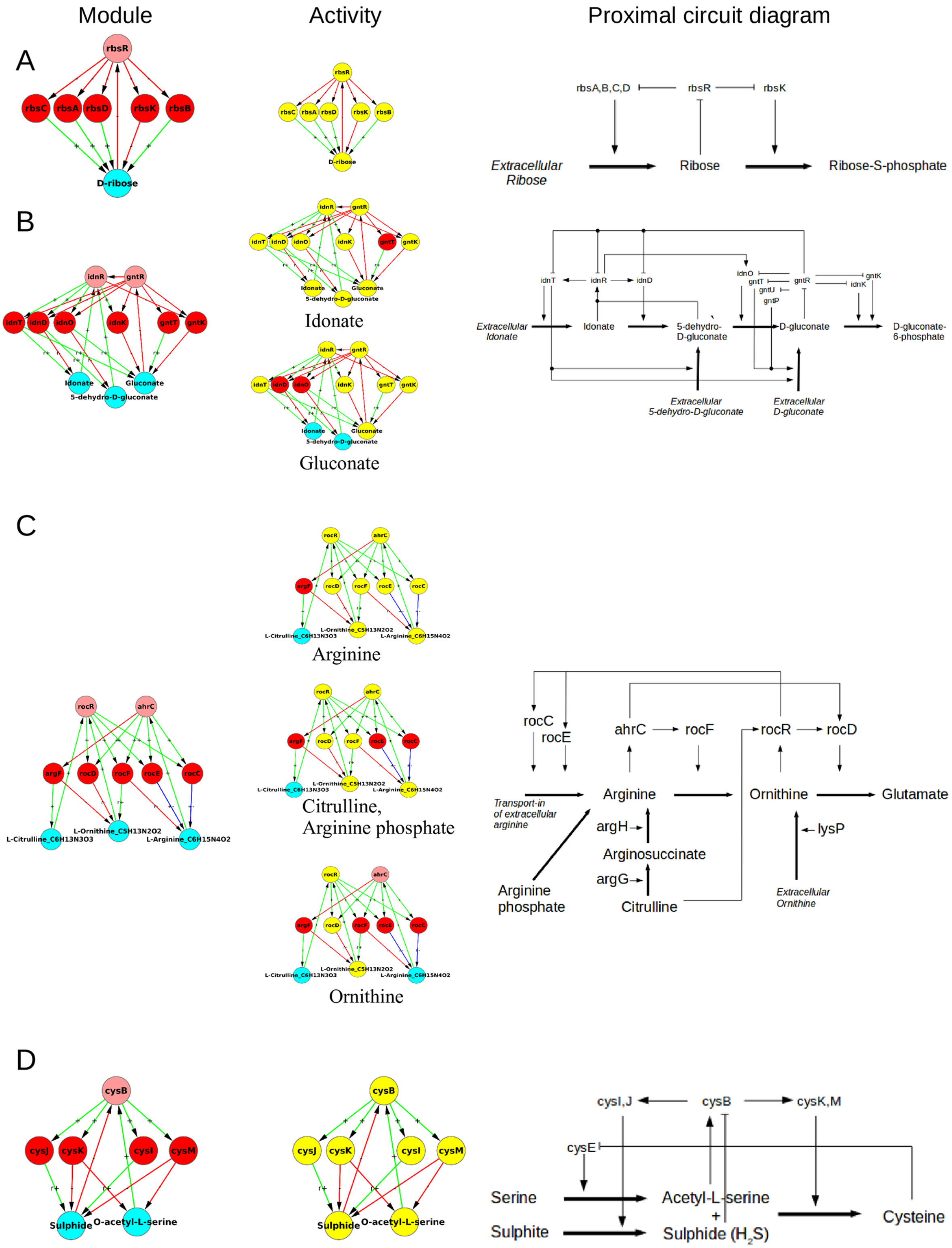


Fig 11. Modules of higher size (≥ 5). A gist of modules of size ≥ 5 of various types is shown through examples in the figure. The three columns show the module, the active subgraph (yellow nodes) of the module under the indicated carbon food source, and its proximal circuit diagram. The food source defines a minimal growth medium (environmental condition, EC). By 'active subgraph' we mean the SCC formed when only the reactions essential in that condition are taken into account and the non-essential reactions in that condition are excluded. **(A)** The Ribose module showing an example of a multigene enzyme, a trivial extension of a 4-node consumer motif. **(B)** The

Idonate-Gluconate module showing an example of sequentially nested consumer motifs—a structure that effectively metabolizes multiple food sources located on a single pathway in the metabolic network. **(C)** The Citrulline-Ornithine-Arginine module showing an example of a module whose different subgrahs are active under different conditions. **(D)** The Cysteine module, a new multinode architecture. Nomenclature of nodes is as in Fig 8.

https://doi.org/10.1371/journal.pone.0203311.g011

acetohydroxybutanoate module are both shown in Fig 10B. These modules are active (present in) all the food source growth conditions where they are part of the reaction pathways for synthesis of amino acids, here methionine, isoleucine and valine. The elementary pathway logic for these modules suggests that they perform the familiar NFL functions of homeostasis.

5.2.3 Size $\geq$ 5 modules. As the size of modules increases to 5 and beyond, the complexity of the modules also increases. Several of these modules contain more than one metabolite binding to TFs, and some of them contain both more than one metabolite and TF. They also have 'sub-modules' which are active in different subsets of conditions. Nevertheless, an analysis of their graphs including the sign of the links, their proximal circuit diagram, and conditions in which they are active, allows us to guess their functional roles in most cases without going into further biological details. There are 10 modules of size $\geq$ 5 in *E. coli* graph $\mathcal{G}_C$ and 3 in *B. subtilis*, giving a total of 13 modules. Of these 13, one is a spurious module in *B. subtilis* (that arises because FBA introduces some fictitious reactions in the metabolic network corresponding to biomass production, see detail in S6 Table) and should be ignored. Of the remaining 12 we could identify functional roles quite clearly for 8 (all in *E. coli*), partially for 2 (both in *B. subtilis*) and we were unable to do so for 2 of the modules (both in *E. coli*). Most (9 out of 12) of the modules of size $\geq$ 5 were just more complex versions of the 4-node consumer motifs and the 3-node feedback loops or their combinations. The details are given in S5 and S6 Tables.

A gist of modules of size $\geq$ 5 is shown in Fig 11. Broadly, three categories emerge: (1) Modules where multiple genes contribute to the formation of an enzyme, e.g., the Ribose module, (Fig 11A) and Phenylpropanoate module (given in S5 Table). In the Ribose module (Fig 11A), the genes *rbsA*, *rbsB* and *rbsC* contribute to the formation of the ribose-ABC transporter. The *RbsD* protein functions for efficient uptake of ribose when it is transported into the cell via a mutated form of glucose transporter (*PtsG*). This 7 node motif is thus essentially like a 4-node consumer motif in which 4 of the enzyme coding genes are performing the same function of transporting the metabolite into the cell. (2) Multiple simple motifs joined together, e.g., Idonate-Gluconate module (Fig 11B). This is a set of three mutually reinforcing consumer motifs. There is a metabolic pathway of conversion of idonate to 5-dehydro-D-gluconate to D-gluconate to D-gluconate-6-phosphate in the cell. The first three metabolites can also be external food sources. Inspection of the proximal circuit diagram and the nodes highlighted in yellow in Fig 11B reveals that when D-gluconate is the food source this circuit will switch on the corresponding consumer motif to produce the last molecule in the pathway, while the part of the circuit before D-gluconate is switched off. Similarly, when Idonate is the food source, the enzymes corresponding to its uptake and metabolism are active (highlighted in yellow) while transporters of the other intermediate food sources (D-gluconate, 5-dehydro-D-gluconate) are off. As another example of the situation in which different parts of the module are active in different conditions we show the Arginine-Ornithine-Citrulline module in Fig 11C. This also decomposes into simpler motifs active in different conditions. Some nodes and links of the module are shared by parts active in different conditions; these are thereby multitasking and contributing to the overall economy of the structure. Here again the structure of the different active parts is essentially one that has already been encountered earlier in size 3 and 4 modules. (3) Modules which are qualitatively different in structure and dynamics from those of size 3 and 4. An example of this is the Sulphide-Acetylserine module (or Cysteine module) (Fig 11D), discussed below.

The following logic can be ascribed to the Sulphide-Acetylserine (or Cysteine) module for the controlled production of Cysteine, assuming that Serine and Sulphite are available in the cell. If Cysteine levels are sufficiently high, it inhibits the enzyme *CysE* that catalyzes the conversion of Serine to Acetyl-L-serine, thereby blocking its own production. When Cysteine level falls sufficiently low, the inhibition of *CysE* is lifted, allowing for the production of Acetyl-L-serine. This activates *CysB* which in turn activates *cysI,J* leading to the production of Sulphide from Sulphite. *CysB* also activates *cysK,M* which catalyze the reaction between Acetyl-L-serine and Sulphide (both now available as reactants) to produce Cysteine. Overproduction of Cysteine shuts the module off by inhibition of *CysE*. Sulphide at high level also represses *CysB* to control its own over-production. This module is well known in the biological literature for the regulation of Cysteine production [54]. This supports the claim that our algorithmic and blind approach starting from the databases is capable of retrieving biologically relevant functional modules at the local level in the network, while also providing information about the global organization of the modules in the network.

We emphasize that our procedure of identifying the SCCs is blind to the signs of edges between the nodes. Still the signs in the modules we obtain through our procedure conform very well to the intuitive logical functioning of these modules. This suggests thats our approach which employs construct of SCC is actually able to find out functionally relevant biological modules or dynamical systems.

The two modules that defied a proper classification, namely, the Serine-Pyruvate-Glycine module and the Allantoin-Glyoxylate module are also more complex (both in *E. coli* and active in 3 and 10 conditions respectively; see S5 Table).

6 Structure of core

Consider the largest SCC (LSCC) in *E. coli* graph $\mathcal{G}_C$ (Fig 5) and its location in $\mathcal{G}_C$ (S1 Fig). We can call the LSCC as the 'core' of the network as it contains the largest number of feedbacks and is positioned at the top in the hierarchical structure of the network. This position provides it the capacity to influence the whole of the downstream GRN and thereby potentially cause a global change in the state of the cell when conditions so demand. The LSCC or core of the *E. coli* graph $\mathcal{G}_C$ comprises of a total of 97 nodes (5 σ-factors, 54 TFs, 8 non-coding RNAs, 19 enzyme-genes and 11 metabolites), Fig 5. In case of *B. subtilis* the LSCC is in the intermediate levels (S2 Fig), which might be a consequence of the incompleteness of its network. We therefore discuss the LSCC of *E. coli* only in the remainder of this section.

The core of $\mathcal{G}_C$ for *E. coli*, though much smaller than the core of $\mathcal{G}_B$ (compare SCC 28 of $\mathcal{G}_C$ in Fig 5 with SCC 20 of $\mathcal{G}_B$ in Fig 3A) is still quite large and complex and it is difficult to assign a specific functionality to it. In fact it is obvious from the location of SCC 28 in S1 Fig and the fact that links emanating from it go to almost all downstream nodes that the core of $\mathcal{G}_C$ is involved in regulating almost all functions of the cell. Nevertheless it is worthwhile to ask if the core of $\mathcal{G}_C$ has a further substructure that is meaningful.

In this context we take a cue from the $\mathcal{G}_C$ of *B. subtilis*. We note that in the case of $\mathcal{G}_C$ of *B. subtilis* the σ-factors make separate SCCs that have no metabolite node (SCC 10 and 14 in Fig 7). However, in the case of $\mathcal{G}_C$ of *E. coli* the σ-factors are part of the LSCC (the 5 green nodes in Fig 5) along with several metabolite, enzyme and TF nodes. This difference between the SCCs of the $\mathcal{G}_C$ graphs of *E. coli* and *B. subtilis* suggested the possibility that σ-factors in *E. coli* act as a 'glue' for connecting different sub-structures of the 'core'. Proceeding on this lead we asked what would happen to the structure of LSCC of graph $\mathcal{G}_C$ of *E. coli* if we

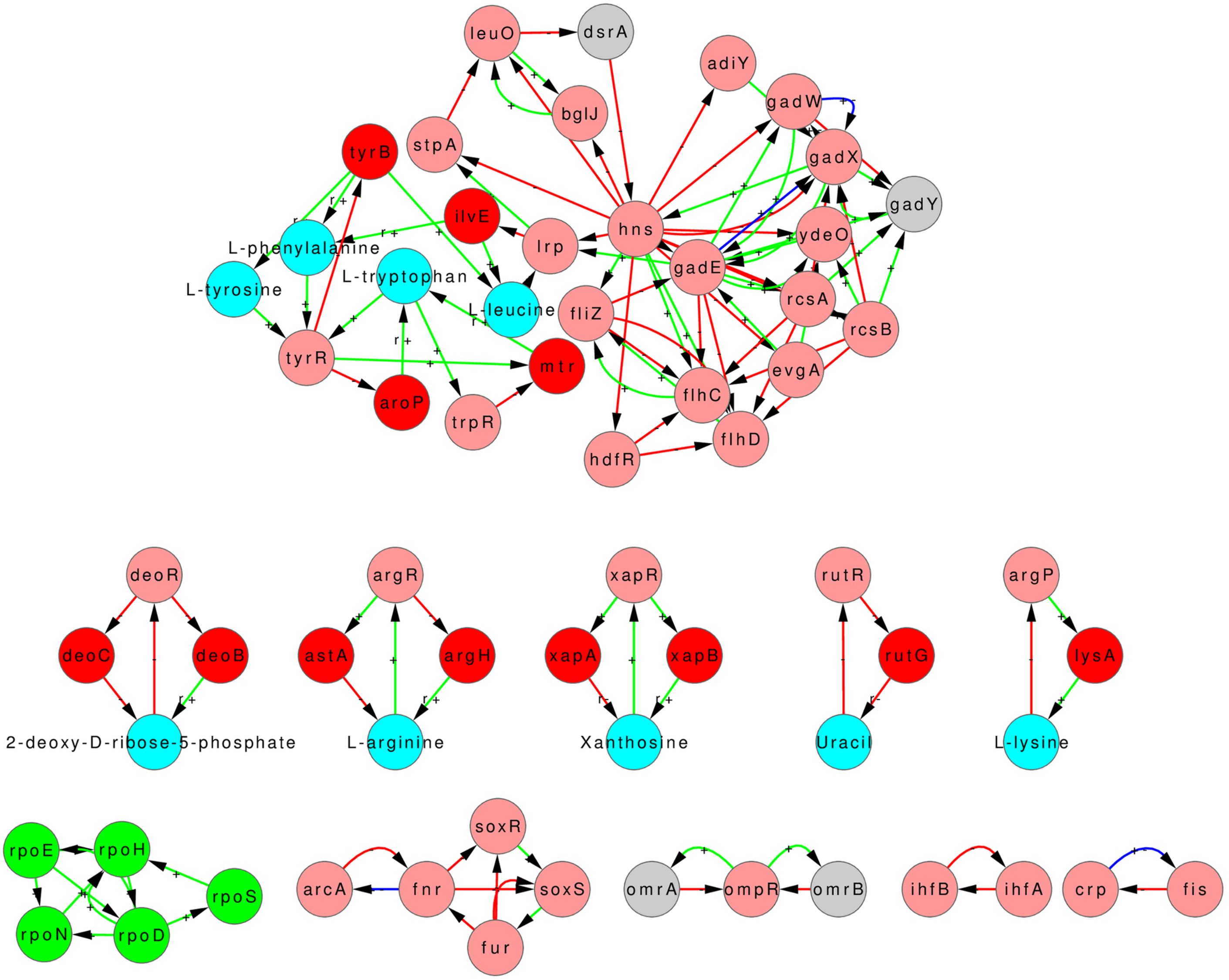


Fig 12. Sub-structure of core of *E. coli* graph $\mathcal{G}_C$. The SCCs constituting the sub-structure of the core revealed upon removal of in-links to σ-factors from the core (shown in Fig 5) of *E. coli* graph $\mathcal{G}_C$.

https://doi.org/10.1371/journal.pone.0203311.g012

separated the SCC formed by the σ-factors from it. This was done by deleting all the incoming links to σ-factors from nodes other than the σ-factors in the graph shown in Fig 5. The number of links deleted was 21. Note that this is a somewhat ad-hoc procedure, but since σ-factors typically have relatively broad roles that play out in good growth condition conditions or different types of stress conditions, this may be justified for our purpose of revealing other close relationships in the network.

Taking out the SCC formed by σ-factors in this manner broke the core into a total of 11 smaller SCCs (including the σ-factor one) the largest of which is of size 29, Fig 12. 5 of the 11 SCCs are composed of only the gene-nodes, while 6 of the SCCs have at-least one metabolite node. The 6 SCCs containing the metabolites can again be classified into modules of size 3 which here are the NFLs (the Lysine and Uracil modules); module of size 4 two of which are consumer-motifs (the 2-deoxy-D-ribose-5-phosphate and Xanthosine modules) and one a combination of two NFLs (the L-arginine module). The largest of the module obtained upon the removal of the SCC formed by σ-factors, size 29, has predominantly genes coding for TF and involve multiple global regulators with varying functions, which prohibits assignment of a simple function to this module. The functioning of the core remains an open question.

7 Discussion

In this work we have: (i) described the causal structure of GRNs without and with feedbacks from MN, (ii) presented a framework and automated method using graph theory and FBA for determining modules in organisms from a knowledge of their GRN, MN and feedbacks from the MN to GRN, and (iii) applied the method in (ii) to *E. coli* and *B. subtilis* to produce a list of modules in these organisms and discussed the functional role of each module.

7.1 Causal and computational structures of GRNs

We employed graph theoretic procedures of strongly connected components (SCCs) and leaf-removal to identify cyclic regions of the GRN and to organize it into a hierarchical structure composed of different levels. The hierarchical organization of the GRNs in section 2 shows that the complicated looking GRNs of *E. coli* and *B. subtilis* can be organized into a causal, largely acyclic architecture, with a few islands of feedback, Fig 1B, consistent with earlier studies. The largest SCC sits near the top of the hierarchy, and is significantly larger and has many more feedbacks than the second largest SCC.

Further, our work takes into account the feedback from metabolic network into the gene regulatory network while illustrating its causal and computational structure, Fig 4, S1 and S2 Figs. We find that most feedbacks from the MN to the GRN affect the network only locally while a few feedbacks have a global impact, Fig 2B and 2C. Further, some of the qualitative features of the organization (hierarchical structure with mostly small islands of feedback and a large globally regulating SCC near the top of the hierarchy) do not change substantially—witness the similarity of the condensed graphs of $\mathcal{G}_{\mathcal{B}}$ in Fig 4 to those of $\mathcal{G}_{\mathcal{A}}$ in Fig 1B. However, the feedback from the MN substantially increases the complexity of the GRNs of both *E. coli* and *B. subtilis* by increasing the number and sizes of SCCs, especially the size of the largest SCC (Figs 3, 5, 6 and 7) and introducing additional levels in the condensed graphs, Fig 4, S1 and S2 Figs. Further, certain enzymes that were at the lowest level in the hierarchy earlier, along with metabolites move up in the hierarchy above many TFs, thereby changing the causal ordering of the hierarchy.

7.2 Automated method of determining modules

We have also obtained condition specific feedbacks from the MN to the GRN by identifying essential reactions in minimal media environmental conditions using FBA. We have used the augmented GRN so obtained, denoted $\mathcal{G}_{\mathcal{C}}$, to find modules in the combined GRN-MN. This procedure can be automated once the following three inputs are available: (1) the GRN of an organism, (2) its MN along with an FBA model, and (3) a database of metabolite-TF interactions in the organism. These inputs are likely to be accessible for more and more organisms. Then the above method can be used to automatically find modules in the organism. The procedure does not find all modules but those that it does seem to have a fairly tight functional role in the organism.

7.3 Module functionality and sign of constituent links

Our method of finding modules by identifying the strongly connected components of $\mathcal{G}_C$ does not use information about the sign of the links (positive or negative) but only the directionality of the links. However in assigning functionality, the sign of the links is crucial. E.g., in Fig 10A the consumer motif functionality can be realized only via the two shown cases. If the sign of one link had been different, no simple functionality could have been inferred. For example, had the sign of the link from *gutR* to *gutP* been negative, the consumer motif functionality would not have been realized. Similarly, in other modules, the combination of signs of links from the source to the destination node is crucial for understanding module functionality. For a given directed subgraph, only a few combinations of the signs of links endow the subgraph with a simple (easy to identify) functionality. The fact that our construction, blind as it is to the sign of the links, finds subgraphs where the signs happen to be just right for assigning a simple functionality to the subgraph, is a non-trivial validation of the approach. It should be mentioned, however, that a module with a 'wrong' sign might have a function that is more complex (e.g., a dynamical property that permits a better response to a time varying signal). Identifying such functionalities requires a more detailed analysis that is beyond the scope of the present work. It is interesting that most modules have a sign combination that allows for an easy-to-interpret functionality; modules that are difficult to interpret functionally are rare (though there are a few, see S5 Table).

7.4 Modules of *E. coli* and *B. subtilis*

Each module represents a dynamical system in its own right. The circuit diagrams we have given represent a starting point for constructing Boolean or ordinary differential equation based models for these modules. We also provided evidence in section 6 that the core or LSCC of *E. coli* had its own internal modular structure.

7.5 Limitations and future directions

7.5.1 Robustness of the condensed graph and its biological interpretation. A question arises as to whether the SCCs and the condensed graph we have constructed after including metabolic feedbacks are robust to future network change as databases expand. We have shown that almost all the SCCs other than the core have an identifiable functionality in the organism. In view of the fairly clear biological role we have identified for them, it is unlikely that future development of the databases will cause these SCCs to disappear, because the role these SCCs can perform in contributing to the organisms' fitness has been identified. We do expect that many of the SCCs will be enhanced to include other nodes and links in the future. However this assertion cannot be made at this point about the core which controls the organism globally, and about the hierarchy in the new condensed graph we have obtained. We do not as yet have a very compelling logic for why some of the parts that are present in the core should appear there or why the hierarchical structure of the condensed graph should be what it is. This is a topic to be investigated in the future.

7.5.2 Extensions of the approach. **(a)** In our determination of modules of *E. coli* and *B. subtilis*, we included the feedbacks from MN. In the process we considered only a subset of the MN. That subset was composed of the essential reactions of the metabolic network. Taking only the part of the metabolic model that is constituted by the essential reactions is somewhat restrictive. It may be useful to go beyond the essential reactions in order to get better modules. We have tried using a flux vector which is obtained as a solution of the FBA (and contains more than the essential reactions) to obtain condition specific graphs. This in fact yields additional modules and also adds additional nodes and links to some existing modules. However since flux vectors are not unique (while essential reactions are) a systematization of this approach is required for further extension. **(b)** Similarly, as discussed at the end of the penultimate para in section 5.2.1 in the context of the D-xylose module the inclusion of next to nearest neighbors in the metabolic graph will also lead to some enlarged modules. **(c)** Enzyme inhibition: The GRN of an organism includes genes coding for enzymes that catalyze the metabolic reactions. In our hierarchical picture these genes come in the level-0 (initiated schematically in red in Fig 2B). The metabolites of the MN can bind allosterically to enzymes (in addition to TFs) and alter their activity. The effect of this on the GRN would be comparatively local because the genes that code for enzymes belong to level-0 in our hierarchical picture of the GRNs. An example of such feedbacks is present in the tryptophan system in *E. coli* which has been well studied and mathematical models have been developed for it (see [55, 56] and references therein). Some enzymes, e.g., those located at the beginning of a metabolic pathway, can have long distance impact on the metabolic network. Our work does not include such feedbacks. We expect that such feedbacks would enrich the modules already obtained by bringing in elements of the metabolic network and thus provide a more complete picture of the modules. (d) Protein-protein interactions and protein degradation: The protein-protein interaction (PPI) network is yet another important cellular network. Our work here is limited in scope in trying to understand the structure of GRNs in the light of MNs. A possible extension of our methodology could consider inclusion of PPIs [10, 36]. Similarly the degradation of a downstream protein causes a feedback upstream [57]; such feedbacks have also been excluded from our analysis. (e) Alarmones: Even though we include protein-ligand interactions, a certain class of small molecules called alarmones that become functional during harsh environmental conditions are not included in this work due to their absence in the parent data sets. An example is the molecule ppGpp, Guanosine tetraphosphate, which regulates the growth of the cell when there is a scarcity of amino acids and during other stress conditions [58, 59]. Our analysis is amenable to the inclusion of such information and it can be done with the availability of this information in future versions of the data sets.

7.5.3 Dynamics of the cell. Our work here is concerned with the structure of bacterial GRNs. A natural extension to it is the understanding of the dynamics through mathematical modelling and simulation. A number of approaches exist for modelling GRNs [60], including Boolean and ordinary differential equation based methods. In any mathematical modelling of a dynamical system one builds upon an understanding of its structure. An insight about dynamical modelling of GRNs is provided by our approach: first individually model the SCCs of the GRN which have feedbacks and are thus more complex, then integrate them in the causal flow of the hierarchy. It may be useful to pursue this 'divide and rule' strategy for a large system such as the GRN.

8 Methods

8.1 Hierarchical organization of nodes

We describe the procedure used in the paper to arrange the nodes of a network into hierarchical levels. The algorithm is as follows: (1) Graph condensation. This involves identification of the SCCs of the graph, and substituting a single representative node for each SCC in place of its constituent nodes. This renders the graph acyclic. (2) Successive identification, assignment and removal of leaves (nodes with no out-degree). In the condensed graph, the leaves identified in the first iteration are identified as level 0 nodes. They are removed. The new leaves identified in the second iteration are identified as level 1 nodes. This process is repeated until all the nodes in the graph are assigned to a level. We demonstrate this using a toy network. This procedure is a modification/combination based on works by Jothi et al [13] and Yu et al [7]. Fig 13 illustrates the procedure.

Panel **A** of Fig 13 shows the toy graph. It is a directed network consisting of 10 nodes and 12 edges. Panel **B** indicates the non-trivial SCC in the same network by enclosing nodes '2', '3', '4' and '5' of the network in a red dashed ellipse. Panel **C** shows the directed acyclic graph (DAG) obtained after the condensation of the toy network. The nodes forming the SCC have been clubbed into one node represented by the bigger circle. Panel **D** shows the nodes of the condensed graph, i.e., the obtained DAG, arranged in a tree layout. The network in panel **D** can be used to assign nodes to different hierarchical levels.

Panel **E** illustrates levels of the hierarchical organization. The categorization of the nodes of the network into finer levels employs iterative leaf removal algorithm on the DAG in panel **D**. The leaves of a network are the nodes with no out-degree. The leaves are assigned to a level and then removed from the DAG which modifies the DAG and exposes new leaves. This procedure is iterated until all the nodes of the DAG are exhausted and the nodes assigned to subsequent levels. This is done as follows. In the DAG in panel **D**, the nodes '8', '9' and '10' has no out-degree and are the leaves. We assign these nodes to level 0 (see panel **E**) and remove them along with their in coming links from the DAG. This leaves us with the modified DAG where '7' is the new leaf. Then '7' is assigned to the higher level, 1, and removed. This procedure is iterated assigning '6' to level 2, '2,3,4,5' to level 3 and finally '1' to level 4. The hierarchical levels along with the nodes are shown in panel **E** for the original toy network where we have expanded the SCC into its constituents nodes. In a GRN, level 0 generally contains many nodes and depending upon the need or for purposes of clarity, the level 0 might not have been shown in the hierarchical pictures used in the other sections of the paper.

8.2 GRN with feedback from MN: $\mathcal{G}_{\mathcal{B}}$

We used the information about metabolic reactions inside the cell from publicly available genome scale metabolic models of *E. coli* [40] and *B. subtilis* [41]. The common genes between the GRN and corresponding metabolic model were used to place an edge from genes of the GRN to the metabolites of the reactions catalyzed by corresponding proteins. The metabolic model of *E. coli*, iJR904 [40], contains 904 genes, 931 reactions and 761 metabolites, while the metabolic network of *B. subtilis*, iBsu1103 [41], contains 1103 genes, 1437 reactions and 1381 metabolites. The influence of metabolites on TFs were obtained from RegulonDB [37, 47] and Ecocyc [42] for *E. coli*, while for *B. subtilis* this information was obtained from Goelzer et al [43]. We selected the reactions from the respective metabolic models (*i*JR904 for *E. coli* and *i*Bsu1103 for *B. subtilis*) which were catalyzed by the enzymes whose coding genes were also present in the GRN (i.e, graph $\mathcal{G}_{\mathcal{A}}$), and also whose metabolites altered the activity of the transcription factors. In *E. coli* we found 320 such genes catalyzing reactions involving 66 metabolites which altered the activity of 57 transcription factors. This introduced 462 links from its GRN ($\mathcal{G}_{\mathcal{A}}$) to the MN (from 320 genes to 66 metabolites) and 77 links from the MN back into the GRN (from 66 metabolites to 57 TFs). In *B. subtilis* we found 168 genes catalyzing reactions involving 29 metabolites which altered the activity of 24 TFs. This introduced 416 links from its GRN ($\mathcal{G}_{\mathcal{A}}$) to the MN (from 168 genes to 29 metabolites) and 34 links from the MN back into its GRN (i.e., from 29 metabolites to 24 TFs). We now constructed the augmented GRN graph $\mathcal{G}_{\mathcal{B}}$ from $\mathcal{G}_{\mathcal{A}}$ by adding new nodes corresponding to the TF binding metabolites (66 and 29 in the two organisms) and new directed links to and from these nodes. The incoming links to these nodes were from the enzyme coding genes whose gene product catalyzed the metabolic reactions of these metabolites (462 and 416 links). The outgoing links from these metabolites were to the genes coding for the TFs to which these metabolites bind (77 and 34 links).

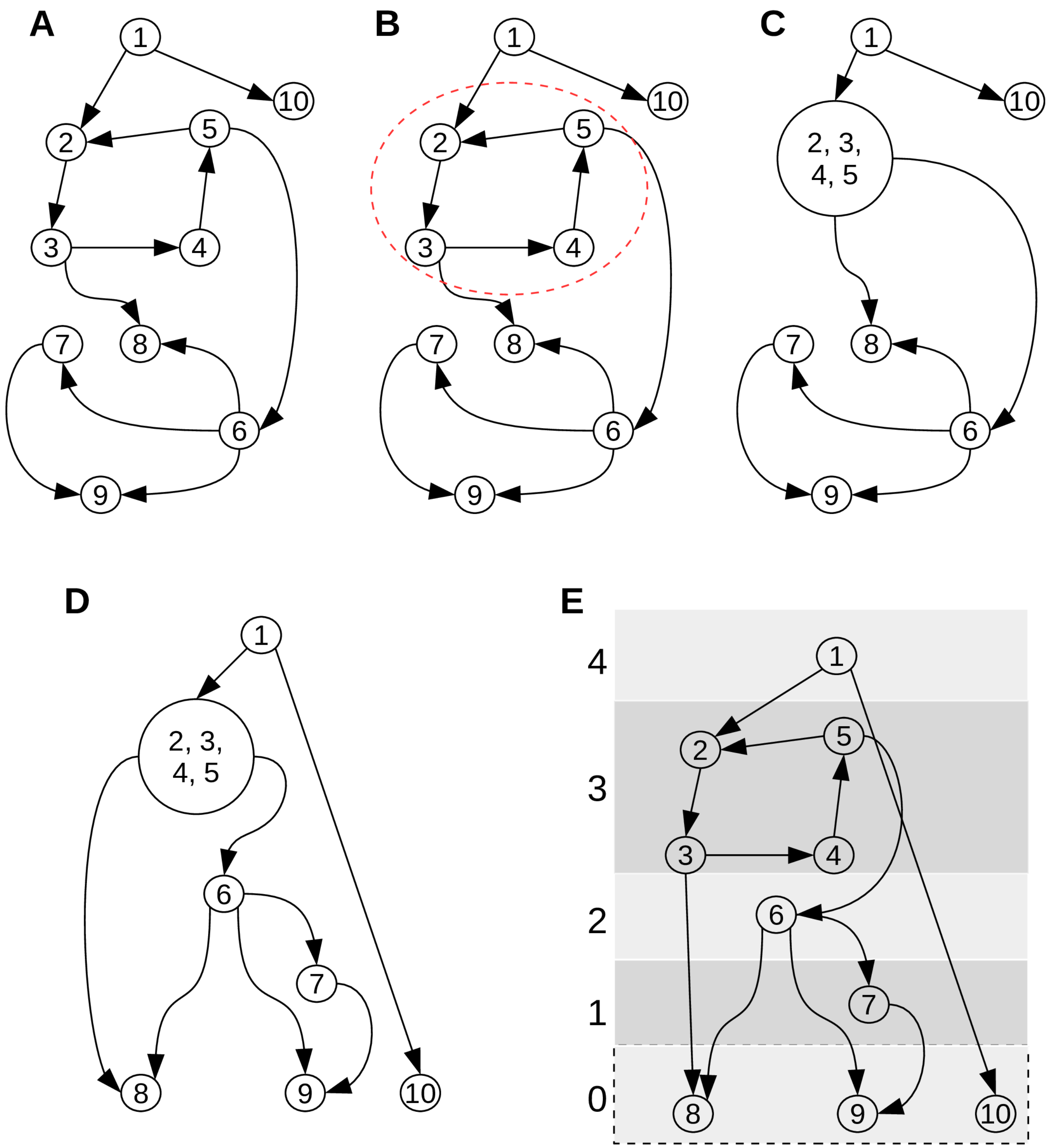


Fig 13. Hierarchical decomposition. Demonstration of arranging the nodes of a toy network into hierarchical organisation used in this paper. The level 0 nodes have been enclosed in a dotted box to indicate that in our figures these nodes have not been shown due to their large number and keep focus on the details of regulation amongst the TFs.

https://doi.org/10.1371/journal.pone.0203311.g013

Thus $\mathcal{G}_{\mathcal{B}}$ contains 66 new nodes and 462+77 new links over $\mathcal{G}_{\mathcal{A}}$ in *E. coli*, and 29 new nodes and 416+34 new links in *B. subtilis*. The metabolite-TF interactions are given in S2 Table. While the metabolic network has substrates (or metabolites) for a number of reactions carried out in the cell for different purposes, we make use of not all but only a subset of the substrates capable of altering the activity of the TFs. Thus, we are using not the whole of the metabolic network but only a part of it (as described below) which directly feeds back into the GRN.

8.3 Construction of graph $\mathcal{G}_{\mathcal{C}}$

8.3.1 Determining the ECs and the corresponding essential reactions. First, we consider the utilization of metabolic pathways by the organism in multiple environmental conditions (ECs). Each EC considered is characterized by a 'minimal' medium in which there is a unique and limited source of carbon together with specified but unlimited other nutrients like nitrogen, phosphorous, sulphur, etc. An EC is characterized as aerobic if external oxygen is available, otherwise anaerobic. We use the computational technique of flux balance analysis (FBA) [44, 51] to determine, for each EC, the reactions of the metabolic network that are essential for the organism to be able to grow in that medium (for details see section 8.4). The FBA simulation in a specified medium gives the maximal steady state growth rate of the organism in the medium and a 'flux vector' that gives the steady state velocity of every reaction in the MN. An EC or medium is said to be viable if the maximal growth rate in it is positive. The model *i*JR904 for *E. coli* has 131 possible carbon sources and *i*Bsu1103 for *B. subtilis* 211 carbon sources. For *E. coli* out of a total of 262 ECs considered (131 aerobic + 131 ananaerobic) only 89 produced viable minimal media in aerobic conditions and 69 in anaerobic conditions. Thus we had a list of 158 viable minimal media for *E. coli*. For *B. subtilis* out of a total of 212 EC (all 212 aerobic, no anaerobic) we had only 118 viable media, all aerobic. The list of viable minimal media is given in S3 Table. We denote the number of viable minimal media so obtained by M ($M = 158$ for *E.coli* and 118 for *B. subtilis*).

8.3.2 Condition specific augmented GRNs and the graph $\mathcal{G}_{\mathcal{C}}$. We then determine the essential reactions of the MN for each of the above ECs. An essential reaction in a metabolic model in a given viable EC, as the name suggests, is one, which, if blocked, results in zero simulated growth (i.e., blocking the reaction changes the EC from being viable to unviable). The determination of essential reactions under a given EC is done in the following standard way. The reaction to be tested for essentiality is first constrained to have a zero flux in the metabolic model, and then the metabolic model is simulated for growth in the particular EC using FBA. If this results in a zero growth value, then the reaction tested is an essential reaction in that EC, else it is not. This procedure is repeated for each reaction in the metabolic model and its essentiality is determined for the given EC. For example, in case of *E. coli* in aerobic glucose minimal growth condition, FBA deemed 218 reactions out of 904 to be essential. Thus for a given EC, in augmenting the GRN with feedbacks from MN, instead of using all 462 links from enzyme coding genes to metabolites we only used the subset corresponding to essential reactions in glucose which were 71 in number. The metabolites that were not participating in the essential reactions were also excluded. We determined the set of essential reactions under each of the 158 ECs for *E. coli* and 118 ECs for *B. subtilis*. The essential reactions for each EC are listed in S3 Table. We augment the GRN with feedbacks from MN using a method similar to that employed previously to arrive at graph $\mathcal{G}_{\mathcal{B}}$, with an additional restriction that the reaction used to link the GRN and the MN must also be an essential reaction under any of the ECs simulated for growth via the metabolic model of bacteria. For each of the ECs we generated a version of GRN augmented with feedbacks from the part of the MN constituted by essential reactions in that EC. This gave us 158 versions of GRNs augmented with condition dependent feedbacks from MN for *E. coli*, and 118 versions for *B. subtilis*. We designate these instances of GRN with condition dependent allosteric feedbacks from metabolic network as graph $\mathcal{G}_{\mathcal{C}_i}$, where the index *i* indicates that the graph is for a given growth condition, labelled by *i*; *i* goes from 1 to *M*. S3 and S2 Tables together contain all the information to construct each $\mathcal{G}_{\mathcal{C}_i}$. We next define the graph $\mathcal{G}_{\mathcal{C}}$ of an organism (*E. coli* or *B. subtilis*) to be union of graph $\mathcal{G}_{\mathcal{C}_i}$ of that organism i.e., $\mathcal{G}_{\mathcal{C}} = \cup_i \mathcal{G}_{\mathcal{C}_i}$. The graph $\mathcal{G}_{\mathcal{C}}$ includes all the nodes and links that are present in any of the $\mathcal{G}_{\mathcal{C}_i}$. Another equivalent way of arriving at the same graph $\mathcal{G}_{\mathcal{C}}$ is to first find the set of essential reactions in the MN for each EC, take the union of the sets of essential reactions across all ECs, find its intersection with the reactions corresponding to the links between enzyme coding genes and metabolites in $\mathcal{G}_{\mathcal{B}}$, and then use the latter set of reactions to augment the GRN with feedbacks from the MN. While the second procedure is a simpler way to arrive at graph $\mathcal{G}_{\mathcal{C}}$, the first procedure presents an opportunity to study the condition specific GRNs augmented with feedbacks from MN individually as well. The number of nodes and edges of various types in graph $\mathcal{G}_{\mathcal{C}}$ are listed in Table 2.

The graph $\mathcal{G}_{\mathcal{C}}$ is a sub-network of graph $\mathcal{G}_{\mathcal{B}}$. All the nodes other than metabolite nodes in $\mathcal{G}_{\mathcal{B}}$ are included in $\mathcal{G}_{\mathcal{C}}$ and their mutual links present in $\mathcal{G}_{\mathcal{B}}$ are also included in $\mathcal{G}_{\mathcal{C}}$. However, a metabolite node in $\mathcal{G}_{\mathcal{B}}$ belongs to $\mathcal{G}_{\mathcal{C}}$ if and only if the metabolite is produced or consumed in a reaction that is essential in any of the *M* media considered. A link from an enzyme node to a metabolite node in $\mathcal{G}_{\mathcal{B}}$ is included in $\mathcal{G}_{\mathcal{C}}$ if and only if the corresponding reaction catalyzed by the enzyme is an essential reaction in any of the *M* media considered.

The advantage of using essential reactions of a metabolic model under an EC lies in their unambiguous determination. An alternative approach (which we have explored but not discussed in detail here) is to use all the reactions with non-zero flux in a flux-vector. It is well known that in a given medium FBA has multiple flux vectors as solutions with the same maximal growth rate and the set of non-zero flux reactions is different for different flux vectors. This gives rise to an ambiguity in the set of reactions to be included if one works with all reactions with non-zero flux. However in a given medium the set of essential reactions is unique for a FBA model.

8.4 Details of the performed FBA

We used the COBRA toolbox in Matlab [61] to perform FBA with metabolic models of *E. coli* [40] and *B. subtilis* [41]. We simulate an environmental condition (EC) by (a) setting the lower and upper bounds of the carbon food source molecule defining the EC to be -10 and 0, respectively, (b) setting the lower and upper bounds of other required nutrients (CO_2, Iron, Water, H^+, Potassium, Na^+, NH_4^+, Phosphate, Sulphate) for growth to be -1000 and 0, respectively, (c) setting lower and upper bounds of other carbon molecules to be 0 and 1000, respectively, (d) setting lower and upper bounds of Oxygen to be -1000 and 1000 (for aerobic medium) or 0 and 1000 for anaerobic medium, respectively, (e) setting lower and upper bounds of ATP maintenance reaction to be 0 and 1000, respectively. The COBRA function `singleRxnDeletion` along with a growth rate cut-off of 1e-6 was used to determine the set of essential reactions for a particular EC. The set of viable food sources, and essential reactions under each EC is given in S3 Table.

Supporting information

S1 Table. $\mathcal{G}_{\mathcal{A}}$, GRN of *E. coli* and *B. subtilis* without feedback from metabolism. The excel sheets list the genes, interactions, regulators, sigma-factors, TFs, ncRNAs, regulated genes,

details of SCCs, and hierarchical levels of the GRNs.
(XLSX)

S2 Table. $\mathcal{G}_B$, GRN of *E. coli* and *B. subtilis* with feedback from metabolism. Lists the genes, metabolites, regulatory interactions, links from the GRN to respective MNs, feedbacks from MN into the respective GRNs, details of SCCs, and hierarchical levels of the GRNs.
(XLSX)

S3 Table. Details regarding FBA of MN of *E. coli* and *B. subtilis*. Contains information about flux balance analysis of the MN of *E. coli* and *B. subtilis*: potential carbon food sources, viable carbon food sources, essential reactions of the organism for each viable minimal medium *i*.
(XLSX)

S4 Table. $\mathcal{G}_C$, GRN of *E. coli* and *B. subtilis* with condition dependent feedback from metabolism. Has sheets related to the GRN augmented with environmental condition dependent feedbacks from MN, $\mathcal{G}_C$: genes, metabolites, regulatory interactions, details of SCCs, and hierarchical levels of the GRNs.
(XLSX)

S5 Table. Regulatory modules of *E. coli*. The functional modules of *E. coli* classified according to their size (number of nodes) accompanied with respective proximal regulatory logic circuit diagrams, the list of environmental conditions in which they are active, and a discussion of their possible functional role.
(XLS)

S6 Table. Regulatory modules of *B. subtilis*. This file contains the same information as in S5 Table, but for *B. subtilis* instead of *E. coli*.
(XLS)

S7 Table. Null model comparision. This file contains the details regarding randomization of a graph to create appropriate null model for comparison of results concerning the number and size of SCCs with those in the original graph.
(PDF)

S1 Fig. The hierarchical structure of GRN of *E. coli*: Condensed version graph $\mathcal{G}_C$. This is a pdf file which contains the figure showing hierarchical structure of graph $\mathcal{G}_C$ of *E. coli*.
(PDF)

S2 Fig. The hierarchical structure of GRN of *B. subtilis*: Condensed version graph $\mathcal{G}_C$. This is a pdf file which contains the figure showing hierarchical structure of graph $\mathcal{G}_C$ of *B. subtilis*.
(PDF)

Acknowledgments

We would like to thank Varun Giri, Parth Pratim Pandey, Srikanth Ravichandran, Areejit Samal and Pooja Sharma for various discussions, and the International Centre for Theoretical Sciences, Bangalore for hospitality where part of this work was done. SK acknowledges a Junior Research Fellowship and Senior Research Fellowship from the University Grants Commission, Government of India. SM acknowledges a Senior Research Fellowship from the Department of Biotechnology, Government of India. SJ acknowledges the Addie and Harold Broitman Membership of the Simons Center for Systems Biology, Institute of Advanced Study, Princeton, USA.

Author Contributions

Conceptualization: Sanjay Jain.

Data curation: Santhust Kumar, Saurabh Mahajan.

Formal analysis: Santhust Kumar, Saurabh Mahajan, Sanjay Jain.

Funding acquisition: Sanjay Jain.

Investigation: Santhust Kumar, Saurabh Mahajan, Sanjay Jain.

Methodology: Santhust Kumar, Saurabh Mahajan, Sanjay Jain.

Software: Santhust Kumar.

Supervision: Sanjay Jain.

Validation: Santhust Kumar, Sanjay Jain.

Visualization: Santhust Kumar.

Writing – original draft: Santhust Kumar, Sanjay Jain.

Writing – review & editing: Santhust Kumar, Saurabh Mahajan, Sanjay Jain.

References

1. Thieffry D, Huerta AM, Pérez-Rueda E, Collado-Vides J. From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli. Bioessays. 1998; 20(5): 433–440. https://doi.org/10.1002/(SICI)1521-1878(199805)20:5<433::AID-BIES10>3.0.CO;2-2 PMID: 9670816
2. Guelzim N, Bottani S, Bourgine P, Képès F. Topological and causal structure of the yeast transcriptional regulatory network. Nature Genetics. 2002; 31(1):60–63. https://doi.org/10.1038/ng873 PMID: 11967534
3. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, et al. Transcriptional Regulatory Networks in Saccharomyces cerevisiae. Science. 2002; 298(5594):799–804. https://doi.org/10.1126/science.1075090 PMID: 12399584
4. Madan Babu M, Teichmann SA. Evolution of transcription factors and the gene regulatory network in Escherichia coli. Nucleic Acids Research. 2003; 31(4):1234–1244. https://doi.org/10.1093/nar/gkg210 PMID: 12582243
5. Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA. Structure and evolution of transcriptional regulatory networks. Current Opinion in Structural Biology. 2004; 14(3):283–291. https://doi.org/10.1016/j.sbi.2004.05.004 PMID: 15193307
6. Ma HW, Buer J, Zeng AP. Hierarchical structure and modules in the Escherichia coli transcriptional regulatory network revealed by a new top-down approach. BMC Bioinformatics. 2004; 5:199. https://doi.org/10.1186/1471-2105-5-199 PMID: 15603590
7. Yu H, Gerstein M. Genomic analysis of the hierarchical structure of regulatory networks. Proceedings of the National Academy of Sciences. 2006; 103(40):14724–14731. https://doi.org/10.1073/pnas.0508637103
8. Farkas I, Wu C, Chennubhotla C, Bahar I, Oltvai Z. Topological basis of signal integration in the transcriptional-regulatory network of the yeast, Saccharomyces cerevisiae. BMC Bioinformatics. 2006; 7(1):478. https://doi.org/10.1186/1471-2105-7-478 PMID: 17069658
9. Lagomarsino MC, Jona P, Bassetti B, Isambert H. Hierarchy and feedback in the evolution of the Escherichia coli transcription network. Proceedings of the National Academy of Sciences. 2007; 104(13):5516–5520. https://doi.org/10.1073/pnas.0609023104
10. Axelsen JB, Krishna S, Sneppen K. The cost and capacity of signaling in the Escherichia coli protein reaction network. Journal of Statistical Mechanics: Theory and Experiment. 2008; 2008(01):P01018. https://doi.org/10.1088/1742-5468/2008/01/P01018
11. Samal A, Jain S. The regulatory network of E. coli metabolism as a Boolean dynamical system exhibits both homeostasis and flexibility of response. BMC Systems Biology. 2008; 2(1):21. https://doi.org/10.1186/1752-0509-2-21 PMID: 18312613

12. Freyre-González JA, Alonso-Pavón JA, Treviño-Quintanilla LG, Collado-Vides J. Functional architecture of Escherichia coli: new insights provided by a natural decomposition approach. Genome Biology. 2008; 9(10):R154. https://doi.org/10.1186/gb-2008-9-10-r154

13. Jothi R, Balaji S, Wuster A, Grochow JA, Gsponer J, Przytycka TM, et al. Genomic analysis reveals a tight link between transcription factor dynamics and regulatory network architecture. Molecular Systems Biology. 2009; 5:294. https://doi.org/10.1038/msb.2009.52 PMID: 19690563

14. Rodríguez-Caso C, Corominas-Murtra B, Solé RV. On the basic computational structure of gene regulatory networks. Molecular BioSystems. 2009; 5(12):1617–1629. https://doi.org/10.1039/b904960f

15. Bhardwaj N, Yan KK, Gerstein MB. Analysis of diverse regulatory networks in a hierarchical context shows consistent tendencies for collaboration in the middle levels. Proceedings of the National Academy of Sciences. 2010; 107(15):6841–6846. https://doi.org/10.1073/pnas.0910867107

16. Corominas-Murtra B, Goñi J, Solé RV, Rodríguez-Caso C. On the origins of hierarchy in complex networks. Proceedings of the National Academy of Sciences. 2013; p. 201300832. https://doi.org/10.1073/pnas.1300832110

17. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature. 1999; 402:C47–C52. https://doi.org/10.1038/35011540 PMID: 10591225

18. Oltvai ZN, Barabási AL. Life's Complexity Pyramid. Science. 2002; 298(5594):763–764.

19. Segal E, Kim SK. The modular era of functional genomics. Genome Biology. 2003; 4:1–2. https://doi.org/10.1186/gb-2003-4-5-317

20. Wagner GP, Pavlicev M, Cheverud JM. The road to modularity. Nature Reviews Genetics. 2007; 8(12): 921–931. https://doi.org/10.1038/nrg2267 PMID: 18007649

21. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA. 1998; 95:14863–14868. https://doi.org/10.1073/pnas.95.25.14863 PMID: 9843981

22. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nature Genetics. 2003; 34(2):166–176. https://doi.org/10.1038/ng1165 PMID: 12740579

23. Ihmels J, Bergmann S, Barkai N. Defining transcription modules using large-scale gene expression data. Bioinformatics. 2004; 20(13):1993–2003. https://doi.org/10.1093/bioinformatics/bth166 PMID: 15044247

24. Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, Barkai N. Revealing modular organization in the yeast transcriptional network. Nature Genetics. 2002; 31(4):370–377. https://doi.org/10.1038/ng941 PMID: 12134151

25. Girvan M, Newman MEJ. Community structure in social and biological networks. Proceedings of the National Academy of Sciences. 2002; 99(12):7821–7826. https://doi.org/10.1073/pnas.122653799

26. Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Physical Review E. 2004; 69(2). https://doi.org/10.1103/PhysRevE.69.026113

27. Ma HW, Zhao XM, Yuan YJ, Zeng AP. Decomposition of metabolic network into functional modules based on the global connectivity structure of reaction graph. Bioinformatics. 2004; 20(12):1870–1876. https://doi.org/10.1093/bioinformatics/bth167 PMID: 15037506

28. Alon U. Network motifs: theory and experimental approaches. Nature Reviews Genetics. 2007; 8(6): 450–461. https://doi.org/10.1038/nrg2102 PMID: 17510665

29. Shen-Orr SS, Milo R, Mangan S, Alon U. Network motifs in the transcriptional regulation network of Escherichia coli. Nature Genetics. 2002; 31(1):64–68. https://doi.org/10.1038/ng881 PMID: 11967538

30. Ihmels J, Levy R, Barkai N. Principles of transcriptional control in the metabolic network of Saccharomyces cerevisiae. Nature Biotechnology. 2004; 22(1):86–92. https://doi.org/10.1038/nbt918 PMID: 14647306

31. Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO. Integrating high-throughput and computational data elucidates bacterial networks. Nature. 2004; 429:92–96. https://doi.org/10.1038/nature02456 PMID: 15129285

32. Herrgard MJ, Lee BS, Portnoy V, Palsson B. Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in Saccharomyces cerevisiae. Genome Research. 2006; 16(5): 627–635. https://doi.org/10.1101/gr.4083206 PMID: 16606697

33. Yeang CH, Vingron M. A joint model of regulatory and metabolic networks. BMC Bioinformatics. 2006; 7(1):332. https://doi.org/10.1186/1471-2105-7-332 PMID: 16820044

34. Shlomi T, Eisenberg Y, Sharan R, Ruppin E. A genome-scale computational study of the interplay between transcriptional regulation and metabolism. Molecular Systems Biology. 2007; 3:101. https://doi.org/10.1038/msb4100141 PMID: 17437026

35. Seshasayee ASN, Fraser GM, Babu MM, Luscombe NM. Principles of transcriptional regulation and evolution of the metabolic system in E. coli. Genome Research. 2009; 19(1):79–91. https://doi.org/10.1101/gr.079715.108 PMID: 18836036

36. Klosik DF, Grimbs A, Bornholdt S, Hutt MT. The interdependent network of gene regulation and metabolism is robust where it needs to be. Nature Communications. 2017; 8(1):534. https://doi.org/10.1038/s41467-017-00587-4 PMID: 28912490

37. Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muniz-Rascado L, Garcia-Sotelo JS, et al. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. Nucleic Acids Research. 2013; 41:D203–D213. https://doi.org/10.1093/nar/gks1201 PMID: 23203884

38. Freyre-Gonzalez J, Manjarrez-Casas A, Merino E, Martinez-Nunez M, Perez-Rueda E, Gutierrez-Rios RM. Lessons from the modular organization of the transcriptional regulatory network of Bacillus subtilis. BMC Systems Biology. 2013; 7(1):127. https://doi.org/10.1186/1752-0509-7-127 PMID: 24237659

39. Sierro N, Makita Y, de Hoon M, Nakai K. DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information. Nucleic Acids Research. 2008; 36: D93–D96. https://doi.org/10.1093/nar/gkm910 PMID: 17962296

40. Reed J, Vo T, Schilling C, Palsson B. An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). Genome Biology. 2003; 4(9):R54. https://doi.org/10.1186/gb-2003-4-9-r54 PMID: 12952533

41. Henry CS, Zinner JF, Cohoon MP, Stevens RL. iBsu1103: a new genome-scale metabolic model of Bacillus subtilis based on SEED annotations. Genome Biology. 2009; 10(6):R69. https://doi.org/10.1186/gb-2009-10-6-r69 PMID: 19555510

42. Keseler IM, Mackie A, Peralta-Gil M, Santos-Zavaleta A, Gama-Castro S, Bonavides-Martinez C, et al. EcoCyc: fusing model organism databases with systems biology. Nucleic Acids Research. 2013; 41: D605–D612. https://doi.org/10.1093/nar/gks1027 PMID: 23143106

43. Goelzer A, Bekkal Brikci F, Martin-Verstraete I, Noirot P, Bessieres P, Aymerich S, et al. Reconstruction and analysis of the genetic and metabolic regulatory networks of the central metabolism of Bacillus subtilis. BMC Systems Biology. 2008; 2:20. https://doi.org/10.1186/1752-0509-2-20 PMID: 18302748

44. Orth JD, Thiele I, Palsson BO. What is flux balance analysis? Nature Biotechnology. 2010; 28(3): 245–248. https://doi.org/10.1038/nbt.1614 PMID: 20212490

45. Kumar S, Vendruscolo M, Singh A, Kumar D, Samal A. Analysis of the hierarchical structure of the B. subtilis transcriptional regulatory network. Molecular BioSystems. 2015; 11(3):930–941. https://doi.org/10.1039/c4mb00298a PMID: 25599335

46. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research. 2003; 13(11): 2498–2504. https://doi.org/10.1101/gr.1239303 PMID: 14597658

47. Salgado H, Santos-Zavaleta A, Gama-Castro S, Millan-Zarate D, Diaz-Peredo E, Sanchez-Solano F, et al. RegulonDB (version 3.2): transcriptional regulation and operon organization in Escherichia coli K-12. Nucleic Acids Research. 2001; 29:72–74. https://doi.org/10.1093/nar/29.1.72 PMID: 11125053

48. Robison K, McGuire AM, Church GM. A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome11Edited by R. Ebright. Journal of Molecular Biology. 1998; 284(2):241–254. https://doi.org/10.1006/jmbi.1998.2160 PMID: 9813115

49. Zheng D, Constantinidou C, Hobman JL, Minchin SD. Identification of the CRP regulon using in vitro and in vivo transcriptional profiling. Nucleic Acids Research. 2004; 32(19):5874–5893. https://doi.org/10.1093/nar/gkh908 PMID: 15520470

50. Grainger DC, Hurd D, Harrison M, Holdstock J, Busby SJW. Studies of the distribution of Escherichia coli cAMP-receptor protein and RNA polymerase along the E. coli chromosome. Proceedings of the National Academy of Sciences of the United States of America. 2005; 102(49):17693–17698. https://doi.org/10.1073/pnas.0506687102 PMID: 16301522

51. Varma A, Palsson BO. Metabolic flux balancing: Basic concepts, scientific and practical use. Biotechnology. 1994; 12:994–998. https://doi.org/10.1038/nbt1094-994

52. Dekel E, Alon U. Optimality and evolutionary tuning of the expression level of a protein. Nature. 2005; 436(7050):588–592. https://doi.org/10.1038/nature03842 PMID: 16049495

53. Krishna S, Semsey S, Sneppen K. Combinatorics of feedback in cellular uptake and metabolism of small molecules. Proceedings of the National Academy of Sciences. 2007; 104(52):20815–20819. https://doi.org/10.1073/pnas.0706231105

54. Kredich NM. The molecular basis for positive regulation of cys promoters in Salmonella typhimurium and Escherichia coli. Molecular Microbiology. 1992; 6(19):2747–2753. https://doi.org/10.1111/j.1365-2958.1992.tb01453.x PMID: 1435253

55. Yanofsky C. RNA-based regulation of genes of tryptophan synthesis and degradation, in bacteria. RNA. 2007; 13(8):1141–1154. https://doi.org/10.1261/rna.620507 PMID: 17601995

56. Santillán M, Zeron ES. Dynamic influence of feedback enzyme inhibition and transcription attenuation on the tryptophan operon response to nutritional shifts. Journal of Theoretical Biology. 2004; 231(2): 287–298. https://doi.org/10.1016/j.jtbi.2004.06.023

57. Cookson NA, Mather WH, Danino T, Mondragon Palomino O, Williams RJ, Tsimring LS, et al. Queueing up for enzymatic processing: correlated signaling through coupled degradation. Molecular Systems Biology. 2011; 7(1):561. https://doi.org/10.1038/msb.2011.94 PMID: 22186735

58. Dalebroux ZD, Swanson MS. ppGpp: magic beyond RNA polymerase. Nature Reviews Microbiology. 2012; 10(3):203–212. https://doi.org/10.1038/nrmicro2720 PMID: 22337166

59. Hauryliuk V, Atkinson GC, Murakami KS, Tenson T, Gerdes K. Recent functional insights into the role of (p)ppGpp in bacterial physiology. Nature Reviews Microbiology. 2015; 13(5):298–309. https://doi.org/10.1038/nrmicro3448 PMID: 25853779

60. Karlebach G, Shamir R. Modelling and analysis of gene regulatory networks. Nature Reviews Molecular Cell Biology. 2008; 9(10):770–780. https://doi.org/10.1038/nrm2503 PMID: 18797474

61. Schellenberger J, Que R, Fleming RMT, Thiele I, Orth JD, Feist AM, et al. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nature Protocols. 2011; 6(9):1290–1307. https://doi.org/10.1038/nprot.2011.308 PMID: 21886097