Supplement

What determines the assembly of transcriptional network motifs in *Escherichia coli*?

Francisco M. Camas and Juan F. Poyatos
Spanish National Biotechnology Centre, Consejo Superior de Investigaciones Científicas (CSIC), 28049 Madrid, Spain.

1 Transcriptional network

This section includes some specifications on the assembled transcriptional regulatory network (TRN) and the quantification of its main attributes (e.g., measurement of auto-regulation in the network):

Heterodimers as single nodes. The IhfA and IhfB constituents of the heterodimer regulator IHF are encoded in two different operons, *thrS-infC-rpmI-rplT-pheMST-ihfA* and *cmk-rpsA-ihfB*, respectively. Because of their similar genomic architecture and regulation (IhfA and IhfB always work as heterodimer and are both under the same regulation), these operons are represented by a single node in the TRN. Similar reasoning applies to the HupA and HupB transcription factors (TFs), components of the heterodimer HU.

Heterodimers as two nodes. The heterodimer RcsAB, whose corresponding operons encode RcsA and RcsB, does not show the previous behavior. RcsB works independently as a homodimer activator (member of the 2-component system RcsC/RcsB). Moreover, RcsAB regulates *rcsA* but not *rcsB*. We thus considered *rcsA* as an autoregulated operon (AO), with the assistance of the protein RcsB, and the operons encoding RcsA and RcsB as two nodes in the TRN.

gntRKU **operon.** We interpreted *gntRKU* as two separated operons (*gntR* and *gntKU*). This prevents the pseudo-autoregulation of *gntKU* by a constitutive GntR, in which GntR would not regulate itself. This also impedes IdnR regulation over *gntKU* (but not over *gntR*) to establish a "pseudo-loop" (or non-dynamical loop) between *gntRKU* and *idnDOTR*.

Transcriptional feedback loops. A recent study documented four transcriptional feedback loops –with more than one component– in *Escherichia coli*'s

transcriptional network [1]. In our TRN only one of these loops remains. Why is this so? One missing loop is the previously mentioned case of non-dynamical loop constituted by *gntRKU*/*idnDOTR*. The other two missing loops appeared when regulations based only on microarray data were considered, and thus they did not occur in our TRN. The only loop that we did find was that constituted by the *marRAB* and *rob* operon pair (Figure S 4).

Feed-forward loop motifs. We identified 232 feed-forward loops (FFLs) in the TRN (Table S1, and Figure 3.A, main text) [1]. In this list there are two instances that should be considered as "pseudo-FFLs". By this we refer to those motifs in which the gene encoding the Y-TF is not part of any transcription unit (TU) regulated by the X-TF (recall that in a FFL $X \rightarrow Y$, $Y \rightarrow Z$, and $X \rightarrow Z$). In both cases, although *arcA* and *pdhR-aceEF-lpdA* are annotated as the X- and Y-elements, respectively, ArcA only regulates the TU constituted by the *lpdA* gene, which is not including the TF acting as putative Y-element of these FFLs, i.e., PdhR [2].

Comparison between Shen-Orr *et al.*, and Camas and Poyatos transcriptional networks. We examined several features of our TRN (CP network) and that assembled in [3] (SO network), where the concept of network motifs was originally introduced. This includes comparisons on 1) network main properties (Table S6), 2) number of AOs (Table S7), 3) FFLs (Table S8), and 4) distribution of operons in the network multilayered structure (Table S9).

2 Main statistical procedures

Autoregulation. We asked two questions related to the distribution of autoregulation in the TRN. First, we examined the distribution of the 64 autoregulated TFs (we did not consider the exclusive autoregulators) with respect to TF sensing specificity. We used a permutation test in which we maintained the number of TFs with and without upstream regulation but randomized the location of the autoregulated TFs. We then measured the number of

[1] Figures S1-S4 show the incoming and outgoing regulations of low/medium regulon-size Y-operons. We showed also the additional links that constitute FFLs.

[2] A list with the 232 FFLs can be found in our website, http://www.cnb.csic.es/∼jpoyatos. The file *all_FFL.txt* contains the X-, Y- and Z-operons listed in the first three columns. We added a fourth column with the FFL class as defined in [2].

autoregulations located among those TFs without external control (i.e. the first layer of the TRN) and compared to the observed value. The presence of autoregulated TFs in this group is smaller than expected ($p = 0.03$, 10000 randomizations). Second, we analyzed how autoregulation correlated with response specificity, i.e., regulon size. We used a permutation test in which we randomized de location of the autoregulations preserving group size of each specificity class (Fig. 1.B, main text). We repeated this protocol in the subsets of TFs with and without upstream control. Thus, only autoregulations inside each subset were randomized (Fig. 1.C-D). Only hubs without upstream control showed a significant enrichment of autoregulated TFs ($p = 0.0086$, 10000 permutations). Alternatively, low regulon-size TFs lacking upstream regulation exhibited a significant low rate of autoregulation ($p = 0.02$, 10000 permutations).

FFLness. We introduced in the main text FFLness ($\mathcal{F}$) as a measure applicable only to TFs with upstream control and regulating ≥ 1 operon(s) –not including autoregulation. For any of these TFs, $\mathcal{F}$ is the ratio of the number of the FFLs in which the TF acts as Y-element, and the maximum number of FFLs that could be potentially assembled with the number of TFs regulating Y (n_{in}) and its regulon size, n_{out} (Fig. 2.A, main text). To examine the significance of the observed measure, we compared with the mean $\mathcal{F}$ obtained in a network null model, controlling for specificity class. Note that i) FFLness is a normalized magnitude that highlights the statistical relevance of the constituted FFLs, i.e., a few FFLs could be easily assembled in a random way by TFs with large regulons, ii) FFLness is almost independent of regulon-size in the null model (Fig. 2.B-D, main text, continuous gray lines), which shows how this magnitude does not exhibit specificity-dependent biases, and iii) the small value of random $\mathcal{F}$ reflects the small number of FFLs that are constituted on average in the random networks (~ 100 *vs.* 232 in the extant network).

When considering the total set of TFs with upstream regulation (Fig 2.D), and comparing with random networks, we found a significantly high $\mathcal{F}$ for all regulon-size classes (low and medium class TFs, $p < 10^{-4}$; hubs, $p = 8 \times 10^{-4}$). FFLness also significantly decayed with regulon size ($p = 0.004$, comparing FFLness of low and high regulon-size TFs, Wilkoxon rank sum test). The use of the alternative class definition discussed in Table S1 showed similar qualitative results. The specific regulatory interactions associated to the computation of $\mathcal{F}$ for low and medium regulon-size TFs

are plotted in Figures S1-S4.

We applied the same protocol to the subsets of autoregulated/non autoregulated TFs (Fig 2.B-C, main text). We found the same qualitative pattern as before. Although the slope of FFLness decay is larger for autoregulated TFs, we did not found a significant difference (see main text for numerical results and their comparison with those considering adjacent regulation).

Significant coregulations by hubs. We counted how many coregulations were established on average by each possible pair of hubs (23 hubs, 253 pairs) in 10000 randomized networks and compared it with those of the extant network. We obtained in this way a set of 253 unadjusted p-values that were corrected for multiple testing as described next.

FDR. We controlled the False Discovery Rate in situations of multiple testing, i.e., when several p-values are calculated simultaneously. We used the following procedure [4]: let $p_1 \leq p_2 \leq \ldots \leq p_m$ be a set of (ordered) unadjusted p-values, the corresponding adjusted p-values are computed as $\tilde{p}_j = min_{k=1,\ldots,m} \left\{ min \left(\frac{m}{k} \, p_k, 1 \right) \right\}, j = 1, \ldots, m$.

Significant SIMs. SIM motifs correspond to TFs exclusively regulating ≥ 3 operons (under the same interaction type). There are 36 TFs that could act potentially as master regulators of positive SIMs in random networks (i.e., each of them regulating $\geq$3 operons –exclusively or not– with positive sign) and 35 TFs as master regulators of negative ones. For each of these TFs we counted how many operons they regulated in a exclusive way in a set of 10000 randomizations and compared this random score with the one observed in *E.coli* (p-values of positive and negative SIMs were adjusted independently).

3 Genomic features of the autoregulated operons

Orientation of genes adjacent to the TRN operons. Divergent architectures can promote the coregulation of the flanking operons through the shared regulatory region (Fig. 2.E, main text). In particular, when the regulation is exerted by a TF encoded in one of these operons, neighbor regulation and autoregulation are readily associated [5]- [8]. To complement the discussions on this issue in the main text, we asked to what extent this divergent architecture has been selected.

We analyzed the relative orientation of the upstream adjacent gene to each of the 681 operons part of the network (Table S2). Note that such adjacent genes could not be constituents of the TRN. We compared this behavior with a null score (randomizing operon orientations in *E.coli*'s genome, 10000 times, while keeping fixed the number of operons encoded in each strand).

Divergent orientations are particularly observed (Table S2). This bias is stronger in the subset of autoregulated operons and was not observed among non-autoregulated ones. We analyzed this significant signal and observed that it was only found (and further enhanced) in the subset of autoregulated operons without upstream control (Table S2, $\not\rightarrow\circlearrowright$). Note that the orientation of adjacent and *downstream* genes did not show any special bias.

Operon structure. We examined the polycistronic/monocistronic architecture of those AOs that, being part of the low regulon-size class, did not regulate an adjacent operon. While there is no particular bias to either design in those AOs without upstream regulation (3 monocistrons + 3 polycistrons, Table S3), polycistronic AOs are considerable enriched in those under this external regulatory control (3 monocistrons + 12 polycistrons, Table S4) [3] . Thus, autoregulated TFs with low regulon-size and upstream regulation are linked both to the polycistronic design and to the assembly of FFLs. These two architectures have in common a dual logic (global TF + specific TF) acting over a set of genes (Figure 3.B-C, main text).

4 Hierarchical FFLs

The "central unit" was defined in the main text as the set constituted by the operon encoding the TF acting as Y and, when applicable, by those of its Z-operons adjacently located (which included adjacent but also second neighbors)[4]. This definition applies to all low regulon-size TFs with upstream regulation (34 operons, Figures S1-S2) and two additional operons (*nagBACD* and *malT*) –both regulating one adjacent operon and four nonadjacent ones, see comments in Table S1. 28 operons of this set are involved, as Y-elements, in the assembly of 74 FFLs (Fig. 3.A, main text). In addition, 53 different operons act as Z-elements of these FFLs.

[3] Among polycistrons associated to TFs with upstream control those with low regulon size are on average the simplest in terms of TUs, even when they are large (Tables S4-S5).

[4] An example of such central unit is the pair of divergent operons plotted in Fig. 3.B, main text.

Approximately half of the previous Y-operons (15/28) regulate at least one non-adjacent Z-operon (second neighbors excluded, see Table S10 and the Appendix of this supplement). There exist 30 of such nonadjacent *Z*s (nad*Z*s), involving 28 different operons (with two cases of shared nad*Z*s: *galETKM* and *manXYZ*, acting as Z-operons of two different Y-operons). Finally, note that to identify homology, we compared amino-acid sequences by Blast with an E-value threshold of 10^{-10} (other threshold values did not change qualitatively our results).

Central unit - nad*Z*s homology. We searched for those nad*Z*s that encode at least one gene homolog to those of the central unit. We obtained 7 out of 30 nad*Z*s with such relationship (Table S10 and Appendix). This number is bigger than expected by chance ($p < 0.0001$ by randomly reassigning 10000 times the set of all nad*Z*s to the set of central units with the restrictions that i) the number of nad*Z*s regulated by the Y-operon of each central unit is fixed, ii) an operon is never assigned to itself [5], and iii) an operon is never assigned twice to the same block because of the mentioned shared *Z*s).

X-Y homology. There are 15 Y-operons regulating nad*Z*s which constitute 42 different (X,Y) pairs with their respective X-operons. We analyzed the homology between genes encoding the X and Y TFs, respectively. We found 6 cases of homolog pairs (Table S10 and Appendix), larger than expected by chance ($p = 0.0003$, by permuting 10000 times TFs and controlling the cases where an operon is paired with itself).[6]

FFLs without homologies. About two thirds (25/40) of the hierarchical FFLs constituted with nad*Z*s cannot be explained by homology-based models (Figure 3.A, main text). We observed that these nad*Z*s are enriched by operons only encoding transport related proteins, and that they are under the control of CRP. These transporters are functionally related to those transporters encoded in the corresponding central unit, yet they are not homologs. Is the transporter located in the central operon, and thus physically linked to the TF, anyway different to those placed nonadjacently? Homologies across transporters associated to different FFLs groups –a given central unit and its associated nad*Z*s– allowed us to compare aspects of function and genomic location. Examples of these homologies are the MFS-symporters or

[5] This could be possible because in the extant network the operon *malT* is both a Y-operon –with four different *Z*s– and a Z-operon of the Y-operon *dgsA*.

[6] *gadW* is found in both X and Y roles.

ABC-transporters of arabinose and galactose (Fig. 4, main text), and also the glucose and the (very related) N-acetyl-D-glucosamine PTS uptake systems. We found equivalent functions encoded in adjacent or nonadjacent locations. We reported in the main text the comparison between the MFS- and ABC-transporters in the arabinose and galactose systems. Additionally, unlike the glucose uptake system located in nad*Z*s, one of the specific components of a N-acetyl-D-glucosamine PTS transporter is encoded in the central unit (see Appendix).

Hierarchical FFLs *vs.* **polycistronic strategies.** We proposed in the main text how an adaptive model based on the establishment of a hierarchical logic on a small set of genes acts as a unifying determinant leading to the occurrence of both hierarchical FFLs and low regulon-size polycistrons with upstream control (Figs. 3.B-C –main text– and Tables S3-S4).

What aspects could influence the presence of either control strategy in a given context? Reasons for the separation in different operons of coregulated genes than act together in a metabolic pathway has been discussed [9]. In brief, this separation allows differential regulation of each operon (enabling temporal programs of gene expression). A polycistronic architecture might not be considered, in this sense, an optimal solution as it could induce the production of some proteins –encoded in the polycistron– before needed. However, this latter strategy can favor the transference of the encoded enzymatic tools across species by horizontal gene transfer (HGT). Neighbor regulation appears in this context as an intermediate solution, combining differential control and capability for successful lateral transfer [7] . Indeed, a large frequency of these events have been recently reported for neighbor regulators [10].

A prediction of the differential expression model [9] is that genes are arranged such that those encoded on the same operon do not skip functional steps in the pathway. This is precisely what we found for genes distributed among the operons in the central unit and the nad*Z*s (see Appendix). Note however that this result could also be due to the mechanisms explaining how bacterial metabolic networks grow, i.e., by HGT uptake of genes encoding products involved in peripheral reactions [11]. This correlates with

[7] The architecture of divergently transcribed operons also reduces the cost of maintenance and replication of an additional promoter region.

the enrichment of nad$Z$s with genes associated to the first steps of peripheral metabolisms [8].

Genome distance between the central unit and the nad$Z$s. For each central unit, we computed the mean distance to its nad$Z$s and then averaged over all units. We then randomized the full set of nad$Z$s and scored distances as before. The average distance of nad$Z$s to the central unit was not particularly small, even when including second neighbors as nonadjacent operons ($p = 0.1$, 10000 randomizations). We also calculated the "across distance" between the coordinates of each central unit and its associated nad$Z$s with respect to the *oriC* region, as chromosomal periodicity of evolutionarily conserved gene pairs has been also recently discussed [12]. This measure did not show any significant pattern either.

Averaged co-conservation of Y- and Z-operons. We considered the phylogenetic conservation of genes involved in the Y/Z operons through 75 species of γ-proteobacteria. Conservation of a particular gene was determined by reciprocal best-hit with an E-value threshold of 10^{-10} (other threshold values did not change qualitatively our results). We quantified co-conservation of each Y-operon/Z-operon by first averaging the Jaccard index [9] of proximity J for all the possible pairs of genes $(y,\ z)/y \in Y,\ z \in Z$. We then determined the average value of J over the set of 30 pairs constituted by the nad$Z$s with their respective Y-operons, and also for the 10 pairs with adjacent Z-operons (ad$Z$s, including here the second neighbors). The average co-conservation of the pairs $\{Y$, all associated Zs –adjacent or not–$\}$ was significantly larger than expected by randomly reassigning the set of Zs ($p < 10^{-3}$, 10000 permutations) [10] . Moreover, the difference on this averaged co-conservation for nad$Z$s (0.40) and ad$Z$s (0.43) was not significant under the permutation of the ad$Z$/nad$Z$ labels (p=0.32, 10000 times).

[8] The most unquestionable cases of non-neutral evolution among hierarchical FFLs are those constituted with nad$Z$s and which could not be explained by homology-based models (25 cases, Fig. 3.A main text). 19 different operons act as nad$Z$s in these FFLs, from which 12 only encode transport related products –also associated to HGT events [11].

[9] This normalized index is a ratio of the number of species in which both genes coexist divided by the total number of species considered. As a reference, the mean value of J for pairs of genes belonging to the same operon is 0.64 (for this set of Y- and Z-operons).

[10] To avoid that the signal of large co-conservation were only caused by the adjacent Zs, we applied the same randomization protocol only in the set of nad$Z$s. We obtained again that the averaged co-conservation of the pairs $\{Y$, nad$Z$s$\}$ was significantly large (p=0.02, 10000 permutations).

Functional characterization. We examined in the Appendix the functional properties of the proteins encoded in the group of 15 low regulon-size Y-operons regulating nad$Z$s (second neighbors excluded, see also Table S10) and all their associated Z-operons, using EcoCyc database [13]. In some cases, the proteins are members of complexes whose additional constituents are not encoded in this group. We nevertheless enclosed this information in parentheses.

We included a simple cartoon showing the location of these proteins in their associated metabolic pathways. We used arrows or ellipses, crossed by arrows, to denote enzymes and transporters, respectively. When a protein is encoded in the central unit, we colored the corresponding symbol in blue. We used red for proteins encoded in nad$Z$s, and gray for proteins encoded in other operons. Some protein complexes required two colors at the same time.

We also described the previously discussed gene homologies, i.e, those between the central unit and nad$Z$s and those between TFs acting as X- and Y-elements of the FFL. Furthermore, we showed for adjacent regulations the relative direction of transcription with respect to that of the Y-operon: (d), divergent; (u), unidirectional (convergent cases were not found). We also indicated when the ad$Z$ is a second neighbor. Abbreviations: Y-op, Y-operon; nad$Z$, nonadjacent Z-operon; ad$Z$, Z-operon adjacent to the Y-operon (including second neighbors).

References

[1] Cosentino M, Jona P, Bassetti B, Isambert H (2007) *Proc Natl Acad Sci USA* 104:5516–5520.

[2] Mangan S, Alon U (2003) *Proc Natl Acad Sci USA* 100:11980–11985.

[3] Shen-Orr SS, Milo R, Mangan S, Alon U (2002) *Nature Genet* 31:64–68.

[4] Benjamini Y, Hochberg Y (1995) *J R Statist Soc B* 57:289–300.

[5] Kolesov G, Wunderlich Z, Laikova ON, Gelfand MS, Mirny LA (2007) *Proc Natl Acad Sci U S A* 104:13948–13953.

[6] Korbel JO, Jensen LJ, von Mering C, Bork P (2004) *Nat Biotechnol* 22:911–917.

[7] Warren PB, ten Wolde PR (2004) *J Mol Biol* 342:1379–390.

[8] Hershberg R, Yeger-Lotem E, Margalit H (2005) *Trends Genet* 21:138–142.

[9] Zaslaver A, Mayo A, Ronen M, Alon U (2006) *Phys Biol* 3:183–9.

[10] Price MN, Dehal PS, Arkin AP (2008) Horizontal gene transfer and the evolution of transcriptional regulation in *Escherichia coli*. *Genome Biology* 9:R4

[11] Pál C, Papp B, Lercher MJ (2005) *Nat Genet* 37:1372–1375.

[12] Wright MA, Kharchenko P, Church GM, Segrè D (2007) *Proc Natl Acad Sci USA* 104:10559–10564.

[13] Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, Paulsen IT, Peralta-Gil M, Karp PD (2005) *Nucleid Acids Res* 33:D334–D337.

			X-TF												
			LC				MC				HC				
			$\not\to\circlearrowright$	$\not\to\not\emptyset$	$\to\circlearrowright$	$\to\not\emptyset$	$\not\to\circlearrowright$	$\not\to\not\emptyset$	$\to\circlearrowright$	$\to\not\emptyset$	$\not\to\circlearrowright$	$\not\to\not\emptyset$	$\to\circlearrowright$	$\to\not\emptyset$	
			15(16)	30	21(22)	13(14)	5(4)	5	7(6)	4(3)	7	1	9	6	total
Y-TF	LC	$\to\circlearrowright$	0(1)	3	0	1	1(0)	2	1	1	27(31)	0	2	9	47(51)
		$\to\not\emptyset$	0	0	0	0	1	0	0	0	14(18)	0	1	3	19(23)
	MC	$\to\circlearrowright$	0	0	0	1	0	6	0	0	13(9)	0	1	8	29(25)
		$\to\not\emptyset$	0	0	0	0	0	0	0	0	9(5)	0	5	0	14(10)
	HC	$\to\circlearrowright$	1	0	0	0	0	0	1	0	29	0	20	10	61
		$\to\not\emptyset$	0	0	0	0	0	2	0	0	53	0	2	5	62
		total	1(2)	3	0	2	2(1)	10	2	1	145	0	31	35	232

Table S1: Classification of the 232 FFLs in the network based on the regulon size of their respective X- and Y-TFs. LC, MC and HC for low-, medium- and high regulon-size classes, respectively. Subgroups are based on the presence/absence of upstream regulation and autoregulation: $\not\to\circlearrowright$, autoregulated TFs without upstream regulation; $\not\to\not\emptyset$, non-autoregulated TFs without upstream regulation; $\to\circlearrowright$, autoregulated TFs with upstream regulation; $\to\not\emptyset$, non-autoregulated TFs with upstream regulation. Small numbers denote number of instances in each subgroup (TFs only regulating their own operon are not considered; Y-elements have upstream regulation by definition). The use of the "central unit" association implies an alternative classification of FFLs based on the number of *nonadjacent* regulated operons. Following this criterion, *exuR*, *nagBACD* and *malT*, all regulating one adjacent operon and four nonadjacent ones, are considered low regulon-size operons. The minor differences introduced by this latter classification –which is the one used in Fig. 3.A, main text– are enclosed in parentheses.

set	N	$\rightarrow\Rightarrow$	$\leftarrow\Rightarrow$	$\Rightarrow\rightarrow$	$\Rightarrow\leftarrow$	p
TRN	681	43.8	56.2	51.0	49.0	0.0015
$\circlearrowleft$	76	36.8	63.2	51.3	48.7	0.02
$\not\circlearrowleft$	59	47.5	52.5	45.8	54.2	0.39
$\not\rightarrow\circlearrowleft$ *low*	18	16.7	83.3	33.3	66.7	0.004
$\rightarrow\circlearrowleft$ *low*	30	36.7	63.3	50.0	50.0	0.10
$\not\circlearrowleft$ *low*	43	48.8	51.2	46.5	53.5	0.50

Table S2: Relative orientation between upstream/downstream adjacent genes ($\rightarrow$) and TRN operons ($\Rightarrow$). Upstream divergent orientation ($\leftarrow\Rightarrow$) is particularly enriched. $\circlearrowleft$, operons encoding an autoregulated TF; $\not\circlearrowleft$, operons encoding a non-autoregulated TF; $\not\rightarrow\circlearrowleft$ *low*, operons encoding an autoregulated low regulon-size TF without upstream regulation; $\rightarrow\circlearrowleft$ *low*, operons encoding an autoregulated low regulon-size TF with upstream regulation; $\not\circlearrowleft$ *low*, operons encoding a non-autoregulated low regulon-size TF (with or without upstream regulation). p, p-value for enrichment of upstream divergent orientation ($\leftarrow\Rightarrow$).

set		AO	Orientation of adj. regulated operon †	Number of nonadjacent regulated op. ‡	Number of promoters in central unit §
LC	**adjacent regulation**	*acrR*	d	0	1/1
		agaR	d	1	1/1
		cusRS	d	0	1/1
		cynR	d	0	1/1
		evgAS	d	1 [1]	2/1
		gcvA	d	1	1/1
		hcaR	d	0	1/1
		ilvY	d	0	1/1
		mngR	d	0	1/1
		pspF	d	1	3/1
		soxR	d	1	1/1
		torR	d	1 [2]	1/1
	poly.	*ada-alkB*	-	2	2
		emrRAB	-	0	1
		qseBC	-	0 [1]	2
	mono.	*lrhA*	-	2	1
		putA	-	0	1
		trpR	-	4	1
MC		*cysB*	-	6 [1]	1
		exuR	u	4	1/1
		iscRSUA	-	6	1
		tyrR	-	7	1
		phoBR	-	9 [1]	1
HC		*argR*	-	10	2
		cpxRA	d	20	1/1
		crp	d	161 [13]	1/1
		fnr	-	85 [7]	1
		lexA-dinF	-	19 [1]	1
		lrp	-	22 [10]	1
		phoPQ	-	19	2

Table S3: Autoregulated operons without upstream regulation. LC, MC and HC for low-, medium- and high regulon-size classes respectively. In LC without adjacent regulation we distinguish the cases of polycistronic and monocistronic AOs. † d, divergent; u, unidirectional. ‡ Regulated second neighbors included. Calculations based only on microarray data enclosed in brackets. § In those cases with adjacent regulation, we showed number of promoters corresponding to the autoregulated and the adjacent operon, respectively.

set	AO	Orientation of adj. regulated operon †	Number of nonadjacent regulated op. ‡	Number of promoters in central unit §
adjacent regulation	*araC*	d	3	1/1
	betIBA	d	0	1/1
	fecIR	u	0	1/1
	galS	u	2	1/1
	glcC	d	0	1/1
	hypABCDE-fhlA	d	3	2/1
	idnDOTR	d	1	1/1
	malI	d	0	1/1
	melR	d	0	1/1
	metR	d	2 [1]	2/1
	prpR	d	0	1/1
	rhaSR	d,c	0	1/1
	uxuR	u	2	1/1
	xylFGHR	d	0	2/1
	zraSR	d	0	1/1
poly.	*chbBCARFG*	-	0	1
	gadAX	-	1 [9]	2
	hipBA	-	0	1
	hyfABCDEFGHIR-focB	-	0	1
	lctPRD (*lldPRD*)	-	0	2
	mdtABCD-baeSR	-	3	1
	mtlADR	-	0	1
	nikABCDER	-	0	2
	pdhR-aceEF-lpdA	-	2	3
	rbsDACBKR	-	0	1
	srlAEBD-gutM-srlR-gutQ	-	0	2
	tdcABCDEFG	-	0	1
mono.	*dgsA* (*mlc*)	-	4	2
	iclR	-	1	1
	nac	-	4 [2]	1

Table S4: Autoregulated operons with upstream regulation and low regulon size. When there is not adjacent regulation we distinguish the cases of polycistronic and monocistronic AOs. † d, divergent; c, convergent; u, unidirectional. In the *rhaSR* case there is adjacent regulation over both the upstream and downstream neighbors. ‡ Regulated second neighbors included. Calculations based only on microarray data enclosed in brackets. § In those cases with adjacent regulation, we showed number of promoters corresponding to the autoregulated and the adjacent operon, respectively.

set	AO	Orientation of adj. regulated operon †	Number of nonadjacent regulated op. ‡	Number of promoters in central unit §
MC	*cytR*	-	8	1
	dnaAN-recF	-	5	8
	gadE	u	5 [8]	3/1
	glnALG	-	5 [7]	3
	nagBACD	d	4	3/1
	oxyR	-	8 [1]	1
	rcsA	-	6 [1]	1
HC	*dusB-fis*	-	54 [8]	1
	fldA-fur	-	31 [4]	4
	fliAZY	u	15	2/1
	hns	-	20 [21]	1
	marRAB	-	15 [1]	1
	purR	-	15 [2]	1
	rpoE-rseABC	-	51	3
	soxS	-	15 [1]	1
	cmk-rpsA-ihfB ¶	-	56 [7]	4
	thrS-infC-rpmI-rplT-pheMST-ihfA ¶			7

Table S5: Autoregulated operons with upstream regulation and belonging to the medium- (MC) and high regulon-size (HC) classes. † d, divergent; u, unidirectional. ‡ Regulated second neighbors included. Calculations based only on microarray data enclosed in brackets. § In those cases with adjacent regulation, we showed number of promoters corresponding to the autoregulated and the adjacent operon, respectively. ¶ *cmk-rpsA-ihfB* and *thrS-infC-rpmI-rplT-pheMST-ihfA*, encoding the two components of the transcription factor IHF, counted as a single node in the network (see the first section of this supplement).

	SO	CP
nodes	423	681
non-autoregulatory interactions	519	1109
TFs	116	135
$\not\rightarrow$	81	66
$\rightarrow$	35	69
$\circlearrowleft$	59 (10)	76 (12)
$\not\rightarrow\circlearrowleft$	35 (5)	30 (3)
$\rightarrow\circlearrowleft$	24 (5)	46 (9)

Table S6: General features of SO and CP networks. $\not\rightarrow$, TFs without upstream regulation; $\rightarrow$, TFs with upstream regulation; $\circlearrowleft$, autoregulated TFs; $\not\rightarrow\circlearrowleft$, autoregulated TFs without upstream regulation; $\rightarrow\circlearrowleft$, autoregulated TFs with upstream regulation. For autoregulators, we detailed the cases of operons encoding a TF that only regulates its own operon (in parentheses).

SO	CP	cases	SO $\not\rightarrow$	SO $\rightarrow$	CP $\not\rightarrow$	CP $\rightarrow$
$\circlearrowleft$	$\circlearrowleft$	50	29	21	20	30
$\circlearrowleft$	$\not\circlearrowleft$	6	3	3	0	2+(4)
$\circlearrowleft$	Abs	3	3	0	-	-
$\not\circlearrowleft$	$\circlearrowleft$	12	9	2+(1)	3	9
Abs	$\circlearrowleft$	14	-	-	7	7

Table S7: Comparison between autoregulated operons in SO and CP networks. An autoregulated operon in the CP network can be autoregulated ($\circlearrowleft$), non-autoregulated ($\not\circlearrowleft$) or absent (Abs) in the SO network, and conversely. We specified those operons with ($\rightarrow$) and without ($\not\rightarrow$) upstream control. Operons appearing in the network only as target operons in parentheses.

	SO	CP
Coh-1	24	66
Coh-2	2	16
Coh-3	4	6
Coh-4	0	9
Inc-1	5	24
Inc-2	0	8
Inc-3	1	2
Inc-4	0	14
Other	6	87
total	42	232

Table S8: Coherent and incoherent FFLs in SO and CP networks (as defined in [2]). Coh: coherent FFLs; Inc: incoherent FFLs, Other: FFLs with at least one dual-type interaction (see also note 2).

	SO network			**CP network**		
layer	operons	⟳	∅	operons	⟳	∅
1	81	35	46	66	30	33
2	233	17	8	177	20	10
3	87	5	3	113	4	3
4	10	2	0	88	6	1
5	12	0	0	65	7	3
6†				94	6	4
7				49	2	2
8				14	1	0
9				15	0	0

Table S9: Distribution of operons per layer in SO and CP networks. We showed explicitely the distribution of autoregulated (⟳) and non-autoregulated TFs (∅). † The two components of the *marRAB-rob* loop are considered to be located both in the 6th layer.

Y-TF	X-TFs	central-unit products	nonadjacent Z-operons products
AraC	CRP	TF, E	1: T; 2: T; 3: T
Cbl	**CysB**	TF	1: P[E, T]
DcuS-DcuR (2c)	Fnr, NarL	2c, E, T	1: NP[E]
DgsA	CRP	TF	1: TF; 2: T; 3: T; 4: T
GadX	CRP, **GadW**, GadE, RpoS	TF, E	1: RP[E, T]
GalS	CRP, **GalR**	TF, T	1: T; 2: P[E]
GlpR	CRP	TF, E, E	1: RP[E]; 2: T, PTAE; 3: NP[T, E], E
HU	CRP	TF	1: P[E]
FhlA	Fnr, IHF, RpoN	TF, E, E	1: RP[TF, E, T]; 2: RP[E]; 3: RP[E, E]
IdnR	CRP, **GntR**	TF, E, E, T	1: RP[E, T]
MalT	CRP	TF, E	1,2: T, T, U; 3: PTAE
BaeS-BaeR (2c)	**CpxA-CpxR** (2c)	2c, T, T	1: T
NagC	CRP	TF, E, T	1: T; 2: NP[TF, E, T]; 3: T
PdhR	CRP, Fnr, ArcA	TF, E	1: NP[TF, E, T]; 2: NP[E]
UxuR	CRP, **ExuR**	TF, E, T	1: NP[E, T]

Table S10: Characterization of low regulon-size Y-TFs establishing FFLs with at least one nad$Z$. First and second columns: Y and X TFs –homolog pairs in bold (two-component systems are also shown). Third and fourth columns: functional characterization of proteins in the central unit and corresponding nad$Z$s labeled with numbers. This also shows the homology relationship –highlighted by same color– between genes in nad$Z$s and those in the associated central unit. Abbreviations: TF, transcriptional factor; 2c, two-component system; E, Enzyme; T, transporter; PTAE, periplasmic transport-associated enzyme; U, uncharacterized protein; NP, near pathway, products acting in regions of the metabolic pathway near those of the central unit; RP: redundant pathway, including proteins which constitute multienzymatic complexes with those encoded in the central unit; P: pathway, sometimes there is no pathway encoded in the central unit, but in the nad$Z$s. See Appendix for further details.

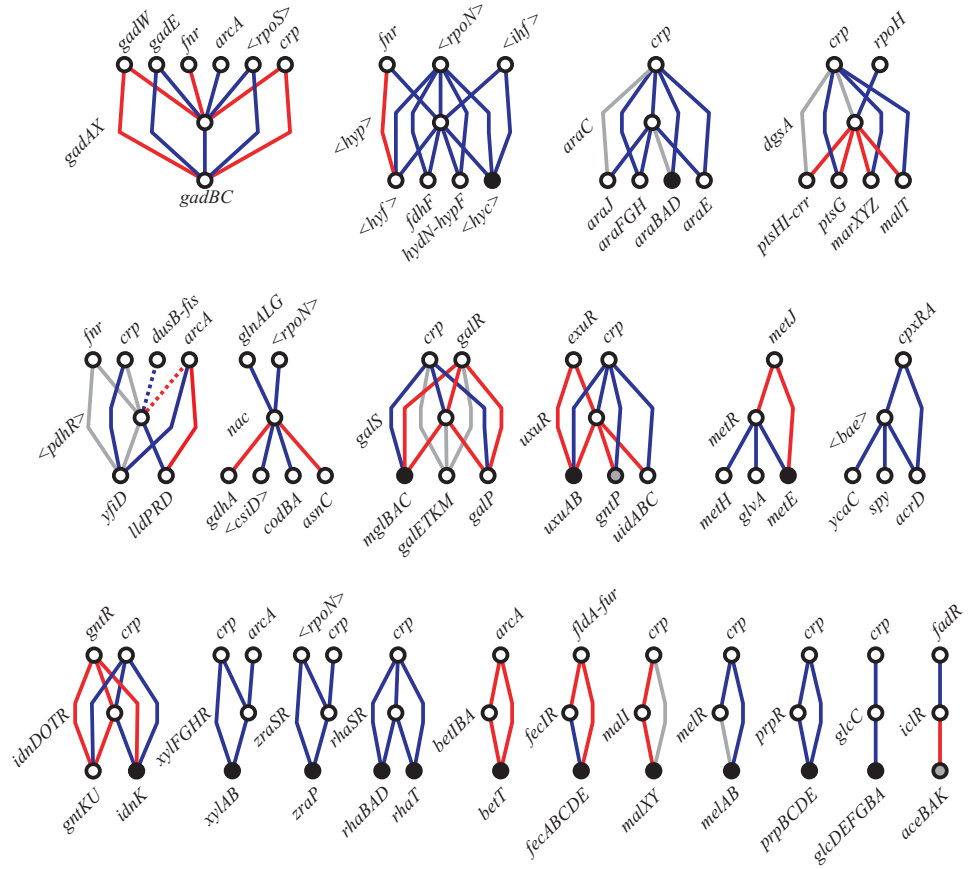


Figure S1: Regulatory links associated to operons with upstream regulation and encoding a low regulon-size autoregulated TF ($1 \leq$out-degree< 5). We showed incoming and outgoing regulations and also those additional ones to describe FFLs (X-Z interactions). Edges color code: blue, activation; red, repression; gray, dual regulation. Z-operons filling color code: black, Z- and Y-operon are adjacent; gray, Z and Y are second neighbors; white, Z and Y are not adjacent. Dashed lines denote links where the TF encoded in the autoregulated operon is not affected by the regulation. This particularly applies to the regulation of *pdhR-aceEF-lpdA* by *arcA*, and leads to the constitution of two pseudo-FFLs. Abbreviations: ⟨*rpoS*⟩, *nlpD-rpoS*; ⟨*hyp*⟩, *hypABCDE-fhlA*; ⟨*hyc*⟩, *hycABCDEFGHI*; ⟨*hyf*⟩, *hyfABCDEFGHIJR-focB*; ⟨*rpoN*⟩, *lptB-rpoN-yhbH-ptsN-yhbJ-npr*; ⟨*ihf* ⟩, *cmk-rpsA-ihfB*; ⟨*csiD*⟩, *csiD-ygaF-gabDTP*; ⟨*bae*⟩, *mdtABCD-baeSR*; ⟨*pdhR*⟩, *pdhR-aceEF-lpdA*; ⟨*srl*⟩, *srlAEBD-gutM-srlR-gutQ*; ⟨*tdcA*⟩, *tdcABCDEFG*. Averaged FFLness: $\langle \mathcal{F} \rangle = 0.64$ (see Fig. 2.B, main text).

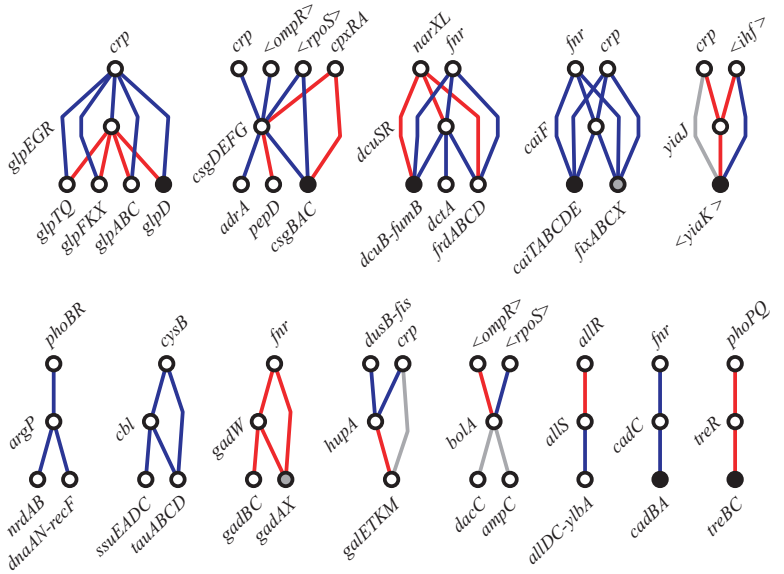


Figure S2: Regulatory links associated to operons with upstream regulation and encoding a low regulon-size non-autoregulated TF ($1 \leq$out-degree< 5). Abbreviations: $\langle$*ompR*$\rangle$, *ompR-envZ*; $\langle$*yiaK*$\rangle$, *yiaKLMNO-lyxK-sgbHUE*, rest of abbreviations as before. Color coding as in Figure S1. $\langle \mathcal{F} \rangle = 0.41$ (see Fig. 2.C, main text).

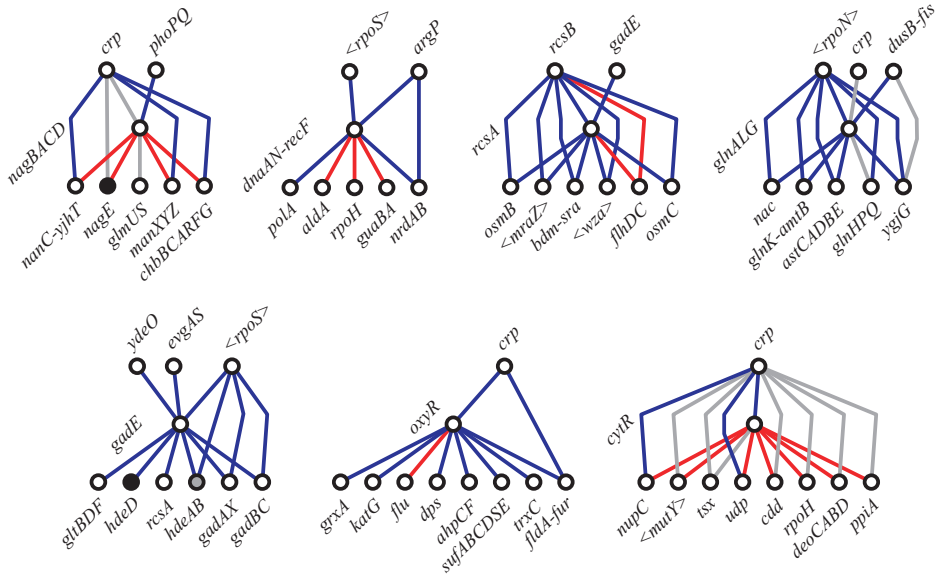


Figure S3: Regulatory links associated to operons with upstream regulation and encoding a medium regulon-size autoregulated TF ($5 \leq$out-degree< 10). In the alternative classification of TFs based on the number of nonadjacent regulated operons *nagBACD* is considered a low regulon-size operon. Maximal FFLness of *rcsA*, *glnALG* and *cytR* corresponds to pairs (X,Y) in which the action of one TF totally relies on the presence of its partner (RcsA on RcsB, RpoN on NtrC –encoded in *glnG*– and CytR on CRP). Abbreviations: $\langle$*mraZ*$\rangle$, *mraZW-ftsLI-murEF-mraY-murD-ftsW-murGC-ddlB-ftsQAZ*; $\langle$*wza*$\rangle$, *wza-wzb-wzc-wcaAB*; $\langle$*mutY*$\rangle$, *mutY-yggX-mltC-nupG*, rest of abbreviations as before. Color coding as in Figure S1. $\langle \mathcal{F} \rangle = 0.38$ (see Fig. 2.B, main text).

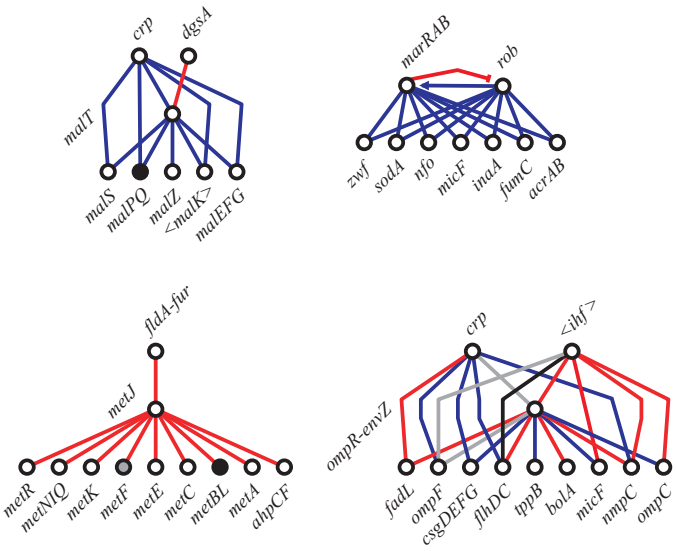


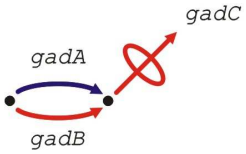
Figure S4: Regulatory links associated to operons with upstream regulation and encoding a medium regulon-size non-autoregulated TF ($5 \leq$out-degree< 10). In the alternative classification of TFs based on the number of nonadjacent regulated operons *malT* is considered a low regulon-size operon. The type of transcriptional interaction between *cmk-rpsA-ihfB* and *flhDC* is not known (in black) . Abbreviations: $\langle$*malK*$\rangle$, *malK-lamB-malM*, rest of abbreviations as before. Color coding as in Figure S1. $\langle \mathcal{F} \rangle = 0.32$ (see Fig. 2.C, main text).

APPENDIX

Autoregulated Y-operon

Y-op: *gadAX*
- *gadA*: enzyme, glutamate dependent acid resistance
- *gadX*: TF

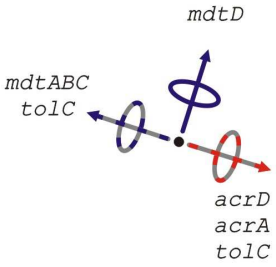
nadZ: *gadBC*
- *gadB*: enzyme, glutamate dependent acid resistance
- *gadC*: APC-transporter (aminobutyrate antiporter)



Notes:
- *gadA* and *gadB* are homologs
- GadX is homolog of the TF encoded in one of its four X-operons, *gadW*. These two operons are second neighbors only separated by the small gene *gadY*

Y-op: *mdtABCD-baeSR*
- *mdtABC* (+ *tolC*): RND-transporter. (drug exporter)
- *mdtD*: MFS-transp. (uncharacterized, drug efflux?)
- *baeSR*: 2-component system

nadZ: *acrD*
- *acrD* (+ *tolC* and *acrA*): RND-transporter (drug exporter)



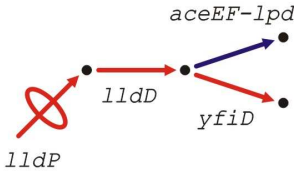
Notes:
- *mdtB*, *mdtC* and *acrD* are homologs
- *baeS* and *baeR* are homologs of the two-component-system genes *cpxA* and *cpxR* respectively; *cpxRA* is the only X-operon for *mdtABCD-baeSR*
- *tolC* encodes the common outer membrane component of several multidrug efflux systems

---

Y-op: *pdhR-aceEF-lpd*
- *pdhR*: TF
- *aceEF-lpd*: pyruvate dehydrogenase

nadZ: *lldPRD*
- *lldP*: LCT-transporter (lactate)
- *lldR*: TF
- *lldD*: lactate dehydrogenase

nadZ: *yfiD*
- *yfiD*: alternative stress induced pyruvate-formate lyase

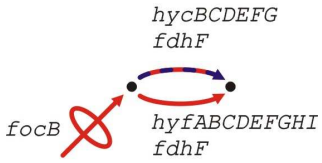


Notes:
- *pdhR* and *lldR* are homologs
- *lldPRD* and *yfiD* are the respective Z-elements of the two pseudo-FFLs (see Fig. S1)
- *lldPRD* is an autoregulated operon

Y-op: *hypABCDE-fhlA*
- *hypABCDE*: proteins for maturation of hydrogenase
- *fhlA*: TF

adZ(d): *hycABCDEFGHI*
- *hycA*: uncharacterized
- *hycBCDEFG*: hydrogenase
- *hycHI*: protein for maturation of hydrogenase

nadZ: *hyfABCDEFGHIJR-focB*
- *hyfABCDEFGHIJ*: hydrogenase (putative)
- *hyfR*: TF
- *focB*: FNT-transporter (formate, putative)



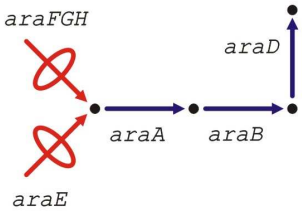
nadZ: *fdhF*
- *fdhF* (+ *hycBCDEFG*): formate-hydrogenlyase complex
- *fdhF* (+ *hyfABCDEFGHIJ*) : formate-hydrogenlyase complex (putative)

nadZ: *hydN-hypF*
- *hydN*: formate-dehydrogenase (putative)
- *hypF*: protein for maturation of hydrogenase

Notes:
- There are multiple homologies between the *hyc* and *hyf* genes
- *fhlA* and *hyfR* are homologs
- *hydN*, *hycB* and *hyfA* are homologs
- *hyfABCDEFGHIJR-focB* is an autoregulated operon

Y-op: *araC*
- *araC*: TF

adZ(d): *araBAD*
- *araBAD*: enzymes in arabinose degradation pathway



nadZ: *araE*
- *araE*: MFS-transporter (arabinose)

nadZ: *araFGH*
- *araFGH*: ABC-transporter (arabinose)

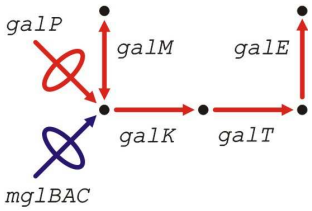
nadZ: *araJ*
- *araJ*: MFS-transporter (uncharacterized, sugar efflux?)

Y-op: *galS*
- *galS*: TF

adZ(u): *mglBAC*
- *mglBAC*: ABC-transporter (galactose)

nadZ: *galP*
- *galP*: MFS-transporter (galactose)

nadZ: *galETKM*
- *galETK*: enzymes for UDP-galactose biosynthesis
- *galM*: galactose-1-epimerase (enzyme that links lactose and galactose metabolisms)



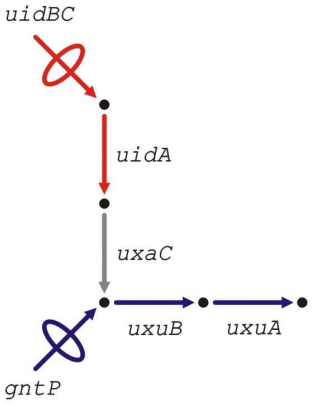
Notes:
- GalS is homolog of the TF encoded in one of its two X-operons, *galR*. The additional X-operon is CRP

Y-op: *uxuR*
- *uxuR*: TF

adZ(u): *uxuAB*
- *uxuAB*: enzymes in fructuronate degradation pathway

adZ(2nd): *gntP*
- *gntP*: GNT-transporter (fructuronate/gluconate)

nadZ: *uidABC*
- *uidA*: enzyme in glucuronide degradation pathway
- *uidB*: GPH-transporter (glucuronide)
- *uidC*: membrane protein associated to *uidB*

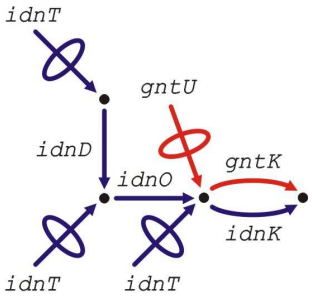


Notes:
- UxuR is homologue of the TF encoded in one of its two X-operons, *exuR*. The additional X-operon is CRP
- *uxuAB* and *gntP* are divergent operons
- *uxaC* is regulated by ExuR. This gene is in the genome neighborhood of *exuR*

Y-op: *idnDOTR*
- *idnDO*: enzymes in idonate degradation pathway
- *indT*: GNT-transporter (idonate/gluconate)
- *idnR*: TF

adZ(d): *idnK*
- *idnK*: enzyme in idonate degradation pathway

nadZ: *gntKU*
- *gntK*: enzyme in idonate degradation pathway
- *gntU*: GNT-transp. (gluconate)



Notes:
- There are multiple homologies between the *idn* and *gnt* genes: *idnT* and *gntU* are homologs and so are *idnK* and *gntK*. Moreover, *idnR* is homolog of the TF encoded in one of its two X-operons, *gntR*, which is located upstream of *gntKU* in the genome. The additional X-operon is CRP

Y-op: *nagBACD*
- *nagBA*: enzymes in N-acetylglucosamine degradation pathway
- *nagC*: TF
- *nagD*: ribonucleotide monophosphatase

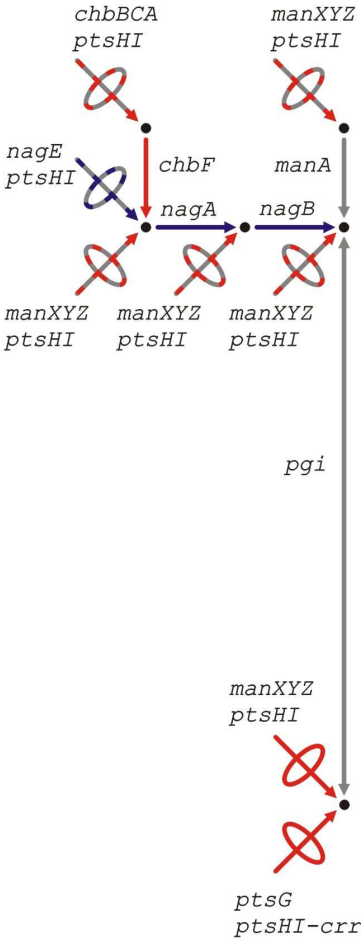
adZ(d): *nagE*
- *nagE* (+ *ptsHI*): PTS-transp. (N-acetylglucosamine)

nadZ: *manXYZ*
- *manXYZ* (+ *ptsHI*) : PTS-transporter (hexoses as N-acetylglucosamine)

nadZ: *chbBCARFG*
- *chbBCA* + (*ptsHI*): PTS-transporter (chitobiose)
- *chbR*: TF
- *chbF*: enzyme in chitobiose degradation pathway
- *chbG*: uncharacterized

nadZ: *nanC-yjhT*
- *nanC*: OmpG-channel(N-acetylneuraminic acid)
- *yjhT*: uncharacterized

Notes:
- *chbBCARFG* is an autoregulated operon
- see Notes for *dgsA* system



Y-op: *dgsA*
- *dgsA*: TF

nadZ: *malT*
- *malT*: TF

nadZ: *manXYZ*
- *manXYZ* (+ *ptsHI*) : PTS-transporter (hexoses as glucose)

nadZ: *ptsG*
- *ptsG* (+ *ptsHI-crr*): PTS-transporter (glucose)

nadZ: *ptsHI-crr*
- *ptsHI-crr*: PTS-transporter (non-specific-sugar components)

Notes:

- *dgsA* and *nagBACD* system (above) are very related: *nagC* and *dgsA* are homologs, and so are *nagE* and *ptsG*; pathways encoded in both systems are closely located in the metabolism and they use the same type of transporters (PTS)

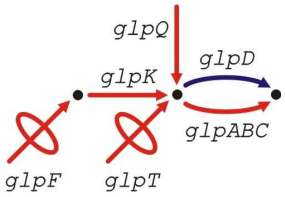
Non-autoregulated Y-operon

Y-op: *glpEGR*
- *glpE*: thiosulfate sulfurtransferase
- *glpG*: intramembrane serine protease
- *glpR*: TF

adZ(d): *glpD*
- *glpD*: glycerol dehydrogenase (aerobic)

nadZ: *glpABC*
- *glpABC*: glycerol dehydrogenase (anaerobic)

nadZ: *glpTQ*
- *glpT*: MFS-transporter (glycerol-3-P)
- *glpQ*: periplasmic transport associated enzyme

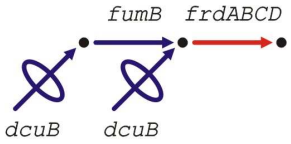


nadZ: *glpFKX*
- *glpF*: MIP-channel (glycerol)
- *glpK*: enzyme for glycerol degradation
- *glpX*: fructose 1,6-bisphosphatase (glycolisis enzyme)

Notes:
- *glpD* and *glpA* are homologs

Y-op: *dcuSR*
- *dcuSR*: 2-component system (anaerobic fumarate respiration)

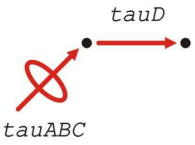
adZ(u): *dcuB-fumB*
- *dcuB*: DCU-transporter (dicarboxylates as fumarate)
- *fumB*: fumarase (anaerobic respiration)



nadZ: *frdABCD*
- *frdABCD*: fumarate reductase (anaerobic respiration)

Y-op: *cbl*
- *cbl*: TF

nadZ: *tauABCD*
- *tauABC*: ABC-transporter (taurine)
- *tauD*: taurine dehydrogenase

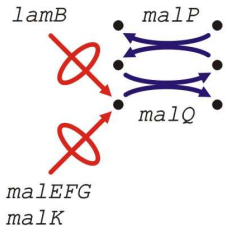


Notes:
- Cbl is homologue of the TF encoded in its only X-operon, *cysB*

Y-op: *malT*
- *malT*: TF

adZ(d): *malPQ*
- *malPQ*: enzymes for maltose and maltodextrins metabolism

nadZ: *malK-lamB-malM*
- *malK* (+ *malEFG*): ABC-transporter (maltose)
- *lamB*: sugar porin (maltose and maltodextrins)
- *malM*: periplasmic protein



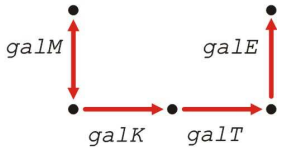
nadZ: *malEFG* (see *malK-lamB-malM*)

nadZ: *malS*
- *malS*: periplasmic maltohexaose transport associated enzyme

Notes:
- *malEFG* and *malK-lamB-malM* are divergent operons: the encoded ABC transporter and porin constitute the maltose/maltodextrin transport system

---

Y-op: *hupA*
- *hupA*: TF



nadZ: *galETKM*
- *galETK*: enzymes for UDP-galactose biosynthesis
- *galM*: galactose-1-epimerase (enzyme that links lactose and galactose metabolisms)