1234–1244 *Nucleic Acids Research, 2003, Vol. 31, No. 4*
DOI: 10.1093/nar/gkg210

Evolution of transcription factors and the gene regulatory network in *Escherichia coli*

M. Madan Babu* and Sarah A. Teichmann

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

Received October 7, 2002; Revised and Accepted December 18, 2002

ABSTRACT

The most detailed information presently available for an organism's transcriptional regulation network is that for the prokaryote *Escherichia coli.* In order to gain insight into the evolution of the *E.coli* regulatory network, we analysed information obtainable for the domains and protein families of the transcription factors and regulated genes. About three-quarters of the 271 transcription factors we identified are two-domain proteins, consisting of a DNA-binding domain along with a regulatory domain. The regulatory domains mainly bind small molecules. Many groups of transcription factors have identical domain architectures, and this implies that roughly three-quarters of the transcription factors have arisen as a consequence of gene duplication. In contrast, there is little evidence of duplication of regulatory regions together with regulated genes or of transcription factors together with regulated genes. Thirty-eight, out of the 121 transcription factors for which one or more regulated genes are known, regulate other transcription factors. This amplification effect, as well as large differences between the numbers of genes directly regulated by transcription factors, means that there are about 10 global regulators which each control many more genes than the other transcription factors.

INTRODUCTION

Regulation of gene expression in an organism involves a complex network. DNA-binding transcription factors are an important component of this network: they respond to changes in the cellular environment by altering the gene expression of relevant genes. Due to this crucial role of transcription factors, they have been studied in many ways, including elucidation of numerous three-dimensional structures. Theoretical analyses of transcription factors in *Escherichia coli* have focused on their sequence families and sequence motifs (1,2). In addition, recent research has elucidated the design principles of the transcriptional regulation network (3–5), including the motifs that recur in the network and their functions.

Our approach is based on the determination of the homology between the domains and protein families of transcription factors and regulated genes, and proteins of known three-dimensional structure. This provides a powerful tool, which goes far beyond sequence comparison methods alone, for finding the domain architecture and evolutionary relationships of transcription factors. Using this method, we can identify uncharacterised *E.coli* proteins that contain DNA-binding domains (DBDs) and identify what is likely to be the large majority of *E.coli* transcription factors.

The homologies between the transcription factors and proteins of known three-dimensional structure yield the domain compositions of the transcription factors. This allows us to quantify the features of this repertoire of transcription factors for *E.coli*: we find that three-quarters of the transcription factors are two-domain proteins, a trend noted previously by Morett and Segovia (6) and Aravind and Koonin (7), and we establish that half of them bind small molecules, a phenomenon first discovered by Jacob and Monod (8).

Based on the domain architectures of the known and predicted transcription factors, we can trace the duplications and recombinations that have produced these proteins in *E.coli* in a more general and extensive manner than has been previously possible (1,7,9). This analysis of domain architecture shows that three-quarters of the transcription factors have arisen by gene duplication.

For the subset of experimentally studied transcription factors, there is information available about the genes they regulate. This allows us to classify these transcription factors in terms of their functions and the numbers of transcription factors and other genes they regulate. We have collated the complete set of transcription factors regulating other transcription factors into a single figure coloured according to the evolutionary family of the DBD, which provides an overview of the central network of gene regulation in *E.coli* at a glance. In order to gain insight into evolution of the entire regulatory network, we have looked for instances of duplications of regulated genes together with their regulatory regions, as well as for duplications of transcription factors together with regulated genes.

MATERIALS AND METHODS

Identification of DNA-binding transcription factors

A preliminary set of transcription factors was identified by extracting all *E.coli* proteins with a DBD. The domains were

*To whom correspondence should be addressed. Tel: +44 1223 402041; Fax: +44 1223 213556; Email: madanm@mrc-lmb.cam.ac.uk

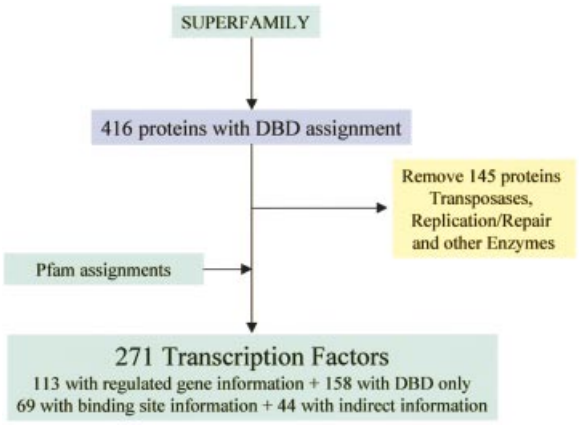


Figure 1. Flow chart of the method used for identification of transcription factors. In addition to our set of 271 transcription factors, there are eight transcription factors without a DBD assignment that have known regulatory information.

identified by the structural annotation system of the SUPERFAMILY database of structural assignments (10,11). The SUPERFAMILY database contains a library of hidden Markov models based on the sequences of domains in the Structural Classification of Proteins (SCOP) database (12,13) and the results of searches by these hidden Markov models against the predicted proteins of completely sequenced genomes. By the assignment of SCOP domains to the *E.coli* proteins, the domain boundaries and family membership of the *E.coli* proteins can be inferred by homology.

In SCOP, evolutionary relationships of domains of known structure are inferred through a combination of clues from sequence, structure and function. Since protein structure is more conserved than sequence in evolution, the structural domain and family definitions in SCOP are much more accurate and extensive than could be achieved by sequence comparisons alone. We refer to the SCOP superfamilies as protein families throughout this work.

The SUPERFAMILY database of structural assignments based on SCOP domains uses hidden Markov models (14), which are probably the most sensitive automatic sequence comparison method currently available (15). The procedure used to make the hidden Markov models and scan them against complete genomes is the iterative SAM-T99 method (16).

As described in Figure 1, SUPERFAMILY assignments were retrieved for the set of 416 proteins in the *E.coli* genome with a DBD assignment. We filtered the set of 416 proteins with DBDs by removing proteins involved in replication/repair, transposases and restriction enzymes according to the functional annotations in GenProtEC (17) and the COGs database (18). We did not include the four σ factors that have structures homologous to the σ^{70} subunit fragment of RNA polymerase (rpoD, rpoS, rpoH, fliA) nor rpoN, which does not have any assigned structure. This resulted in a final set of 271 transcription factors with DBD assignments from the SUPERFAMILY database. In addition to the structural assignments from SUPERFAMILY, 46 of the transcription factors had domain assignments from nine families from the Pfam database (19) of hidden Markov models.

For 121 of the 271 transcription factors, there is experimental information about the genes they regulate in the RegulonDB database (20) and in Shen-Orr *et al*. (4) as well as two references about FIS not in either data set (21,22). For eight of the transcription factors with known regulatory information there were no homologues of known structure detected in the SUPERFAMILY database. Given that 113 out of 121 transcription factors with known regulatory information were assigned a DBD, it is likely that the 271 transcription factors assigned a DBD represent the large majority of all *E.coli* transcription factors. Other calculations have given an upper limit of 400 (3) and 350 (1) transcription factors. Thus our analysis encompasses a sizeable fraction of the entire repertoire of *E.coli* transcription factors and the conclusions we draw based on 271 proteins are likely to hold for the whole set.

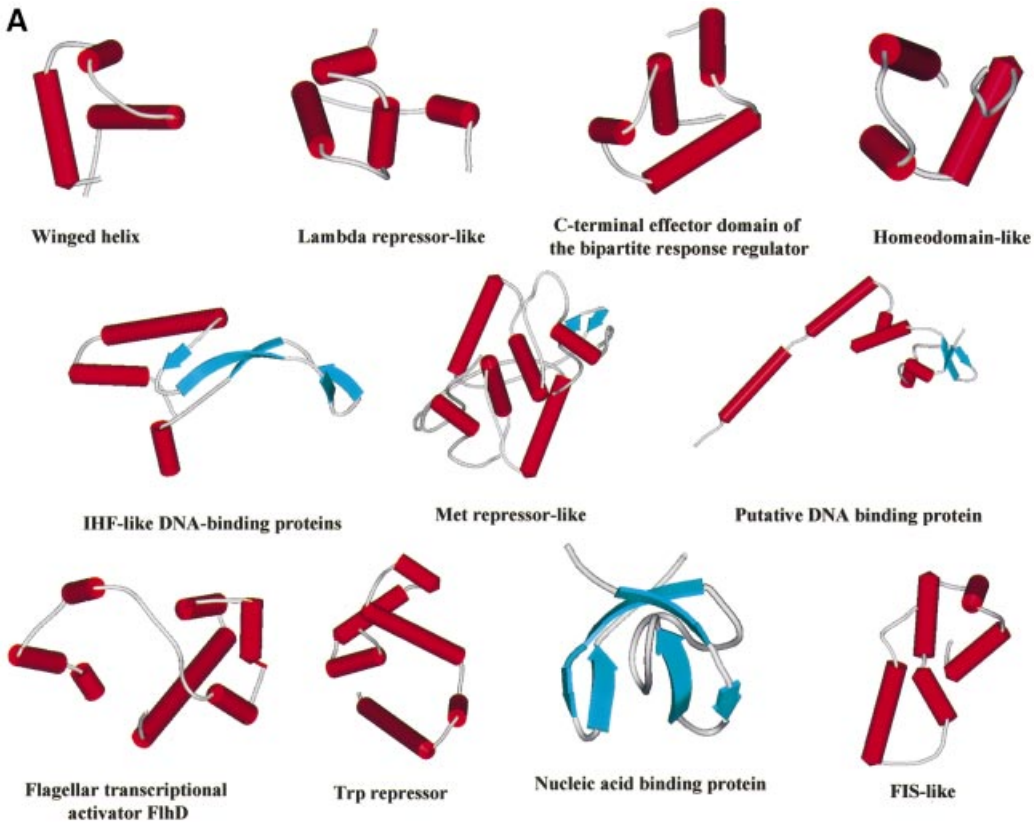
RESULTS

Domains and protein families of *E.coli* transcription factors

Eleven DNA-binding domain families. The domain assignments from the SUPERFAMILY and Pfam databases showed that the 271 transcription factors have DBDs from one of 11 different families. Representative three-dimensional structures of the 11 families are shown in Figure 2A. All families except nucleic acid binding proteins contain a helix–turn–helix motif. In each family, the motif is in an entirely different structural context, as is evident from Figure 2A, and the helix–turn–helix itself is not a feature that necessarily implies evolutionary relationship. In fact, the helix–turn–helix motif also occurs in domains that are not DNA-binding, such as in the enzyme cytochrome c oxidase or the C-terminal domain of ribosomal protein L7 (A. Murzin, personal communication). Therefore, we are adhering to the conservative definition of protein families in the SCOP database and are assuming that the proteins in these different families are not related by descent.

Given that the set of 271 *E.coli* transcription factors have only 11 different DBDs, it is of interest to analyse the distribution of DBDs in these proteins. As shown in Table 1, the sizes of the DBD families vary from 123 members of the winged helix family to a single member in the Trp repressor and nucleic acid binding domain families. The homeodomain-like family is the second largest family with 52 domains; there are two other medium-sized families and the remaining DBD families are small, following a power law type distribution of family size similar to that observed in complete genomes (23,24).

Three-quarters of transcription factors are two-domain proteins. The DBDs generally occur in combination with other domains: there were only 25 single-domain proteins (~10%), but 202 two-domain proteins (~75%), 33 three-domain proteins (~12%) and nine four-domain proteins (~3%). All proteins contain a single DBD except for an uncharacterised protein that contains two copies of winged helix domains and 20 proteins that contain two adjacent homeodomains. There are two separate crystal structures of *E.coli* proteins with two adjacent homeodomains; in one



structure only one of the homeodomains interacts specifically with the major groove and in one of them both domains do so (25,26).

Instead of a second copy of the DBD, the other domain more frequently has a different function, such as a small molecule-binding or enzymatic domain. The set of non-DBDs comes from 46 different families. These 46 families can be classified into five broad functional categories. There are 12 families of enzyme domains, of which at least three are certainly catalytically active: the signal peptidase domain in LexA, the methyl-DNA protein methyltransferase as in Ada and the P-loop nucleotide triphosphate hydrolase in DnaA. For the other nine enzyme domains, it is unclear whether they are catalytically active, as there are known examples where an enzymatic domain has lost its ability to catalyse a reaction and just serves as a small molecule-binding domain (27). There are 18 families of small molecule-binding domains, five protein interaction domain families and 10 domains of unknown function. Finally, there are CheY-like response regulator receiver domains that are phosphorylated by kinases in two-component signal transduction systems. These different categories are indicated by the shapes of the domains in Figure 2B. The large DBD families have partner domains from all of the functional categories and the partner domains can be positioned N- or C-terminal to the DBD, as shown in Table 1.

Overall, small molecule-binding domains are the most frequent type of partner domain, occurring in 44% of the transcription factors. In addition, the Trp and Met repressor DBDs bind small molecules with the same domain that binds DNA. This suggests that almost half of the transcription factors in *E.coli* are directly regulated by the presence or absence of small molecules, as previously noted (27).

The CheY-like response regulator receiver domain occurs in ~10% of the proteins. Protein interaction domains that either interact with RNA polymerase subunits or are involved in dimerization occur in ~7% of the proteins. Enzymatic domains occur in ~5% of the transcription factors. One or two DBDs occur in isolation in ~12% of the proteins. In the remaining the cases, the DBD occurs in combination with a domain of unknown function, or a region for which no known domain assignment can be made.

This distribution of partner domains suggests that only a small fraction of transcription factors are regulated exclusively at the transcriptional level and not by a small molecule or through a sensor protein. The major types of domain combinations are a DBD with a small molecule-binding domain or a Che-Y like response regulator receiver domain with the C-terminal effector domain (25 proteins). There are 120 proteins with 27 distinct combinations of domain families of the DBD with small molecule-binding domain type: the two main domain architectures are winged helix with periplasmic binding protein-like II (43 proteins) and the λ repressor-like with the periplasmic binding protein-like I (14 proteins).

Three-quarters of the E.coli transcription factors have arisen by gene duplication. From Table 1 and the schematic representation of the domain architectures of the transcription factors in Figure 2B, it is obvious that these proteins have

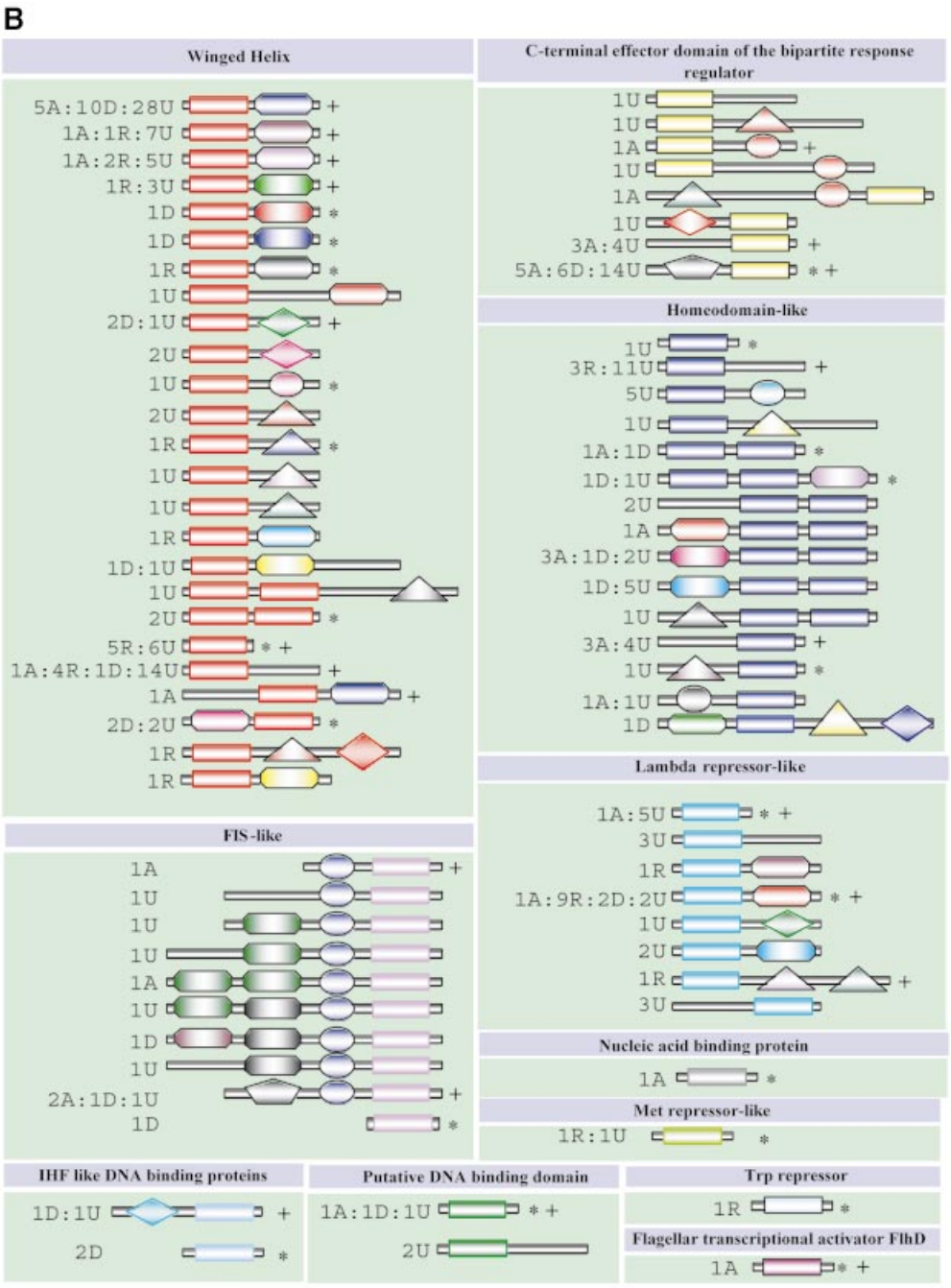


Figure 2. (**A**) (Opposite and above) The three-dimensional structures of the 11 DBD families seen in the 271 identified transcription factors in *E.coli*. The figure highlights the fact that even though the helix–turn–helix motif occurs in all families except the nucleic acid binding family, the scaffolds in which the motif occurs are very different. (**B**) The 74 unique domain architectures of the 271 identified transcription factors. Each functional class is represented by a different shape and each family within the functional class is represented by a different colour. The DBDs are represented as rectangles. The partner domains are represented as hexagons (small molecule-binding domain), triangles (enzyme domains), circles (protein interaction domain), diamonds (domains of unknown function) and the receiver domain has a pentagonal shape. The letters A, R, D and U denote activators, repressors, dual regulators and transcription factors of unknown function, respectively, and the number of transcription factors of each type is given next to each domain architecture. Architectures of known three-dimensional structure are denoted by asterisks, and '+' are cases where the regulatory function of a transcription factor has been inferred by indirect methods, so that the DNA-binding site is not known. The key to this figure, with the name of each family, is available as supplementary data from the website.

evolved by extensive recombination of domains. However, proteins with the same sequential arrangement of domains are likely to be direct duplicates of each other, as discussed in Apic *et al*. (28) and Bashton and Chothia (29). Therefore, we have also looked at whole proteins rather than individual domains and have grouped them into protein families maintaining the same domain architecture. In total, there are the 74 distinct domain architectures shown in Figure 2B, which have duplicated to give rise to 271 transcription factors. Thus 73% of these transcription factors have arisen as a

Table 1. Information about the domain architectures of each DBD family

DBD type	No. of examples	No. of distinct domain architectures	No. of partner families	DBD (N- or C-terminal)
Winged helix	123	25	22	N:20 C:3
Homeodomain-like	52	15	13	N:5 C:8
C-terminal effector domain of the bipartite response regulator	38	8	5	N:3 C:4
λ repressor-like	31	8	6	N:6 C:1
FIS-like	13	10	1	C:9
Putative DBD	5	2	1	N:1
IHF-like DNA-binding proteins	4	2	1	C:1
Met repressor-like	2	1	0	
Nucleic acid binding protein	1	1	0	
Trp repressor	1	1	0	
Flagellar transcriptional activator FlhD	1	1	0	
Total	271	74		

The occurrence, number of distinct domain architectures, number of different partner families and the position of the DBD on the primary sequence is given for each of the 11 DBD families.

Table 2. Regulated genes and functional information for the global regulators (for a complete list please refer to www.mrc-lmb.cam.ac.uk/genomes/madanm/ec_tf/)

C	G	D	I	T	F
CD	mlc	8	10	18	Sugar utilization systems; phosphotransferase system (PTS) and general activator
CD	lrp	54	10	64	Leucine-responsive regulatory protein: amino acid catabolism during carbon starvation
RS	arca	72	3	75	Aerobic respiratory control
RS	narl	65	10	75	Nitrate and nitrite regulation and anaerobic respiration
RS	fnr	112	51	163	Genes in nitrogen metabolism
ES	cspa	2	28	30	Cold shock protein A
ES	crp	197	113	310	cAMP receptor protein and general regulator
IT	fur	21	5	26	Iron regulation and pH sensing
SP	hns	24	5	29	Regulates two fimbrial operons and basic proteins regulator
EH	ihf	100	9	109	Integration host factor; general factor
EH	fis	76	220	296	Factor for inversion stimulation; regulation of rRNA and tRNA operons and other genes

There are 11 global regulators in this table. The columns are: C, functional class; G, gene name; D, direct number of genes regulated; I, indirect number of genes regulated; T, total number of genes regulated; F, function. The individual functional classes are: CD, carbon compound degradation; RS, redox sensing; ES, environment sensors; IT, ion transporters; SP, structural proteins; EH, general enhancer. These global regulators regulate a large number of genes (directly and indirectly through another transcription factor) as opposed to the fine tuners, which regulate a small set of genes, mostly only directly. The complete list of 121 transcription factors, available on the website, contains additional functional classes, namely: CM, carbon compound metabolism; AR, antibiotic resistance; RR, restriction and repair; GP, unclassified (none of the categories).

consequence of complete gene duplication. The protein families maintaining the same domain architecture can contain members with different regulatory activities: there are several domain architectures that are found in both activators and dual regulators or repressors and dual regulators (30).

In the FIS-like DBD family, it is obvious that a two-domain fragment has duplicated with subsequent recombinations with one or two additional domains, so that the DBD forms an evolutionary module with a P-loop-containing nucleotide triphosphate hydrolase. In these proteins, the P-loop domain interacts with the σ^{54} subunit of RNA polymerase. There are three different domain architectures that have a GAF domain N-terminal to these two domains and three other domain architectures in which a PYP-like sensor domain is N-terminal to the two-domain module. This example illustrates how a module of two domains acts as an evolutionary unit that is elaborated to different three-domain modules.

In contrast, there are two examples where a pair of domains is inverted rather than retaining the N- to C-terminal order. Domains of the winged helix family occur both N- and C-terminal to periplasmic binding protein II domains. The C-terminal effector domain of the bipartite response regulator occurs N-terminal to TPR repeat domains in two different architectures and C-terminal in one domain architecture.

It is worth mentioning that the winged helix DBD almost always occurs at the N-terminus, as shown in Table 1. The only exception is the cAMP-binding domain-like family, which occurs N-terminal to the winged helix DBD. The four proteins with this domain architecture are CRP and FNR, which are both global regulators controlling a large number of genes, and two hypothetical proteins.

The organisation of the transcriptional regulatory network

Ten functional categories of transcription factors. The 121 transcription factors for which we have information on their regulated genes can be divided into 10 general functional categories, as shown in Table 2. The largest group of transcription factors, 37 proteins, control genes involved in carbon compound degradation, and another 24 transcription factors control genes in carbon compound metabolism. Twenty transcription factors are redox-sensing proteins that control genes in response to a change in redox status and nine others are environmental sensors for things such as tempera-

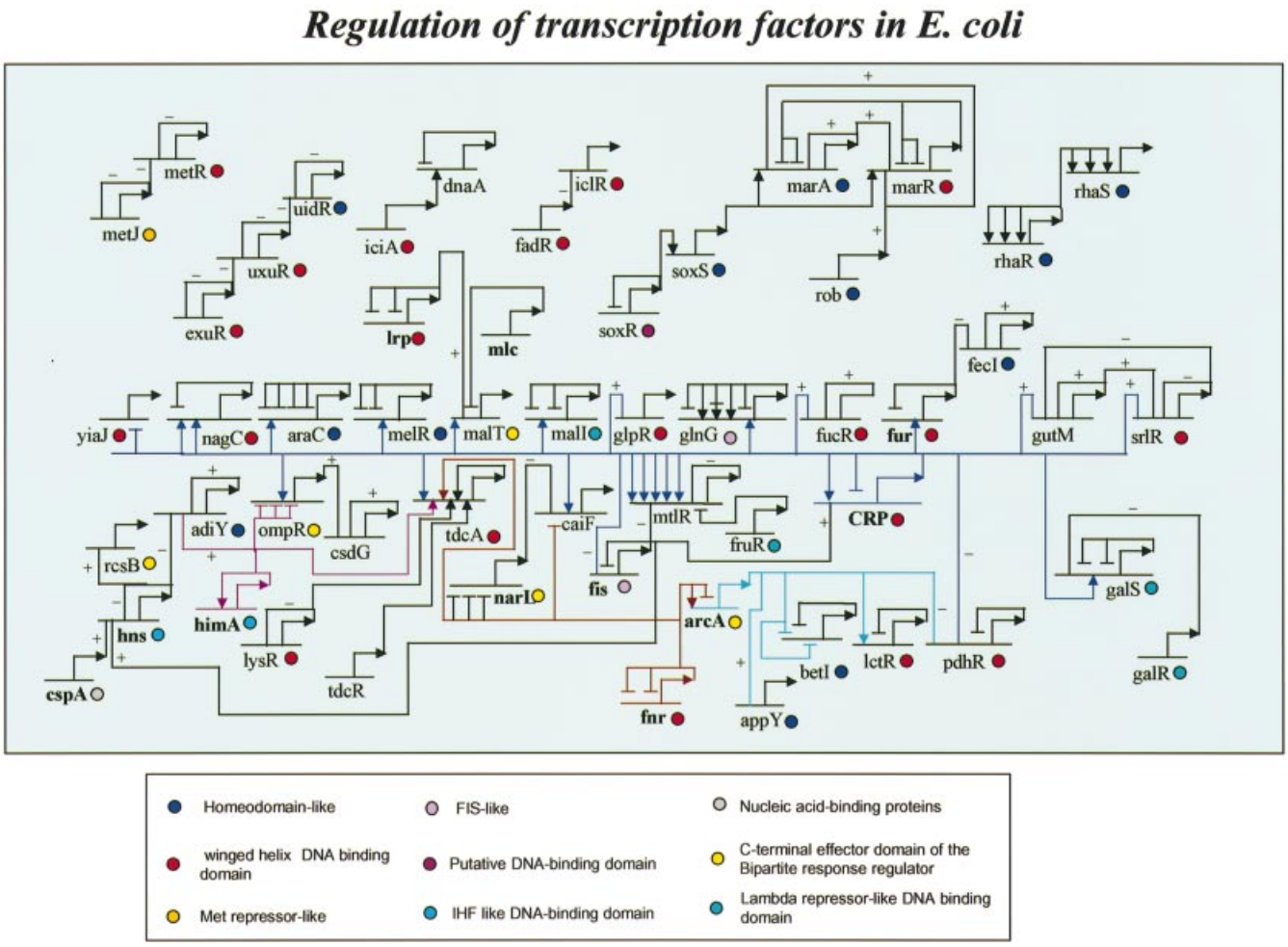


Figure 3. The transcription factor regulatory network in *E.coli*. When more than one transcription factor regulates a gene, the order of their binding sites is as given in the figure. An arrowhead is used to indicate positive regulation when the position of the binding site is known. A horizontal bar is used to indicate negative regulation when the position of the binding site is known. In cases where only the nature of regulation is known, without binding site information, + and – are used to indicate positive and negative regulation, respectively. These examples may be indirect rather than direct regulation. The DBD families are indicated by circles of different colours as given in the key. The names of global regulators are in bold.

ture. Eight proteins control genes involved in antibiotic resistance. Six transcription factors regulate ion transporters and another six control structural proteins. Restriction and repair genes are regulated by six transcription factors. There are two transcription factors that act by bending DNA and thus affecting binding of other transcription factors and the polymerase, which we group in a separate category of general enhancers. Finally, three transcription factors control genes in none of the above categories.

Regulatory cascades: the central part of the transcriptional network. Individually, these 121 transcription factors regulate from 1 to 197 genes and, all together, there are 1302 genes and 303 operons in the regulatory network (transcription factors and regulated genes). For 38 of the transcription factors, some of the regulated genes are themselves transcription factors. There are 34 autoregulatory transcription factors, as listed in Rosenfeld *et al*. (31). To investigate the regulation of transcription factors, we integrated the information available from RegulonDB (20) and Shen-Orr *et al*. (4), as well as Falconi *et al*. (21) and Gonzalez-Gil *et al*. (22), to produce the diagram in Figure 3. This figure shows the network of transcription factors currently known to regulate each other in *E.coli*. CRP controls 18 different transcription factors apart from itself. Two other transcription factors, FNR and ArcA, regulate four transcription factors and FIS and IHF (himA and himD) regulate three transcription factors. CRP is a global sensor of food levels in the environment. FNR and ArcA are involved in sensing the redox status of the cell to regulate genes involved in respiration under aerobic and anaerobic conditions. FIS and IHF-like are general enhancers, which frequently act together with other transcription factors to regulate genes. Thus the transcription factors involved in respiration and growth are those which regulate the most transcription factors.

In fact, there are few long cascades of transcription factors that regulate each other in the *E.coli* gene regulatory network, as noted previously (3,4). In our current dataset as illustrated in Figure 3, there are 23 two-level cascades, 32 three-level cascades and six four-level cascades. Thus even in the simple prokaryote *E.coli*, the transcriptional regulation network is a complex combination of multi-level cascades and motifs.

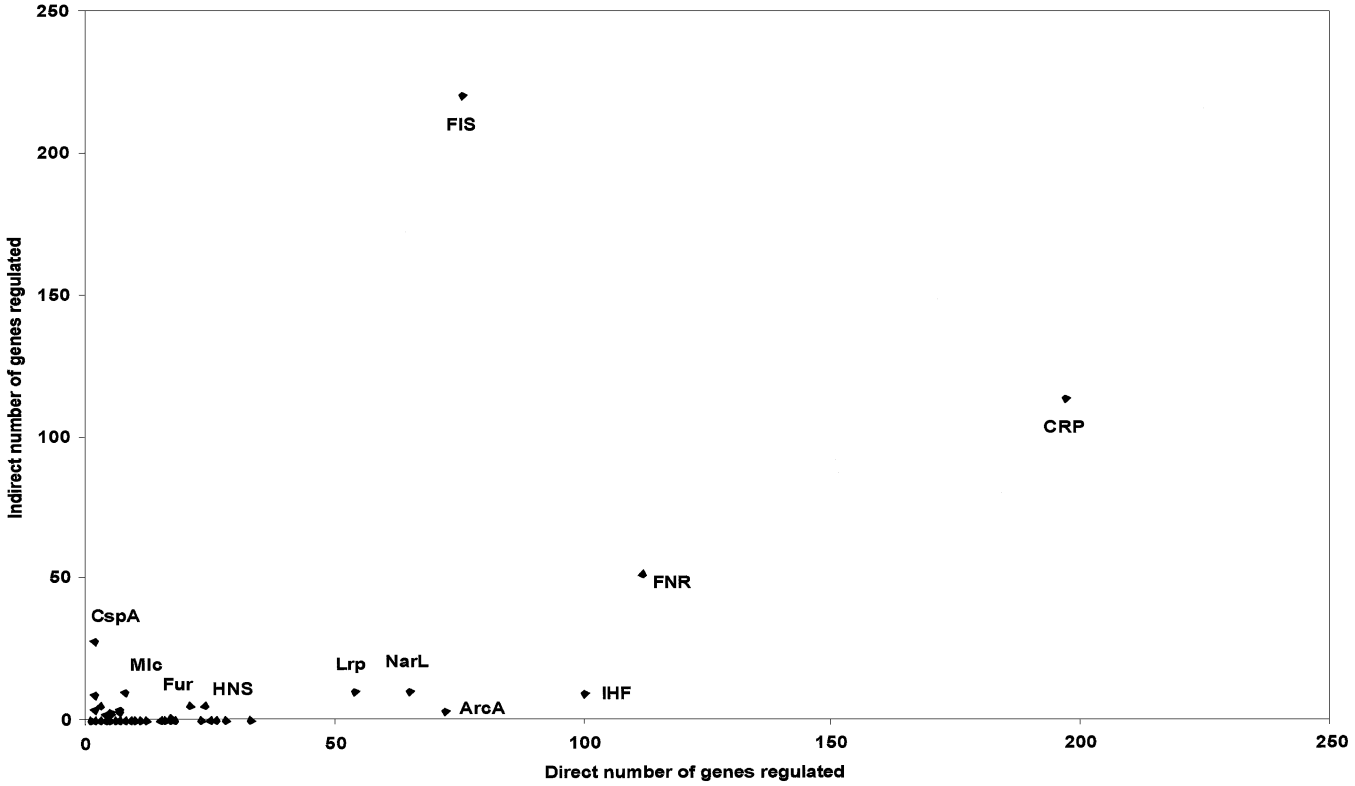


Figure 4. Direct and indirect gene regulation by *E.coli* transcription factors. The direct number of genes regulated is represented on the *x* axis and the indirect number of genes on the *y* axis. The global regulators, which are marked on the graph regulate a large number of genes, and participate in regulatory cascades, resulting in indirect regulation of genes.

Global regulators. Through regulation of other transcription factors, the transcription factors amplify their range of control over genes to encompass a set of indirectly regulated genes. Thus the total number of genes regulated by a transcription factor is the sum of the genes regulated directly and indirectly, given in the third column of Table 2. In 6 of the 10 functional categories, there are transcription factors that regulate genes both directly and indirectly, and control more than 15 genes all together. These 11 transcription factors, shown in Table 2, can be viewed as 'global regulators', as opposed to the remaining transcription factors, which are 'fine tuners'. For a complete list of transcription factors, please refer to the supplementary data at www.mrc-lmb.cam.ac.uk/genomes/madanm/ec_tf/. The difference between these two types of transcription factors is clear from the graph in Figure 4: the global regulators have more directly and indirectly regulated genes than the remaining transcription factors. In a recent analysis of the *E.coli* network motifs by Shen-Orr *et al.* (4), global regulators were defined in a different way than here: as those transcription factors that controlled 10 or more operons. This gives a set of 15 global regulators, nine of which are also in our set of 11 global regulators.

Our set of 11 global regulators are transcription factors involved in carbon degradation (mlc and lrp), redox status sensing (arcA, narL and FNR), ion transport regulation (fur), environmental sensors (cspA and CRP), a regulator of structural proteins (hns) and two general enhancers (IHF and FIS). Thus the global regulators are proteins that control responses to changing food levels and carbon degradation (mlc, lrp, CRP) and transcription factors that respond to changes in redox status or ion levels of the cell (arcA, narL, FNR, fur). Cold shock protein A (cspA) binds RNA and regulates translation in this manner, but it is also known to bind DNA (32). FIS is a homeostatic regulator of general superhelicity.

Eight of the 11 global transcription factors are dual regulators. mlc and fur are only repressors, and cspA is only an activator. cspA is the only global transcription factor that has a DBD of the nucleic acid-binding protein family; the other transcription factors belong to three other DBD families and have seven different domain architectures.

With the current status of experimental data, the remaining 109 transcription factors each regulate 33 or fewer genes in total. Therefore, with the current status of information about the *E.coli* gene regulatory network, it appears that the majority of transcription factors are 'fine tuners' that control a limited, specific set of genes, while a small number of transcription factors are 'global regulators' that control tens or hundreds of genes by direct and indirect influence.

Evolution of the transcriptional regulation network

Regulation by combinations of transcription factors. The organisation of the transcriptional regulation network includes a few global regulatory transcription factors and many fine tuners, regulatory cascades and dense overlapping regulons, in which several transcription factors jointly regulate several operons (4). In our data set compiled from RegulonDB and Shen-Orr *et al.* (4) there is one operon controlled by seven

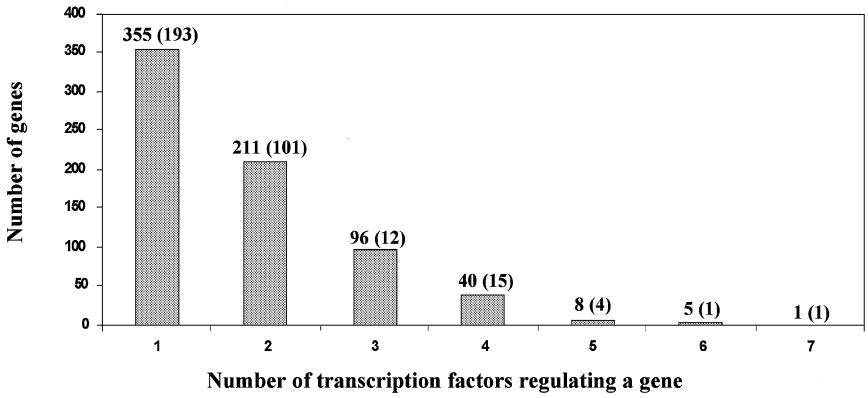


Figure 5. Distribution of the number of transcription factors regulating a gene. Numbers in parentheses represent the number of operons.

Table 3. Numbers of transcription factors active at the same promoter

No. of transcription factors	No. of co-regulating transcription factors
26	0
35	1
18	2
12	3
4	4
6	5
2	6
1	7
2	8
2	9
2	10
1	11
1	12
1	13
1	19
1	20
1	20
1	52

There is one transcription factor that has 52 co-regulating transcription factors (CRP) and 26 transcription factors that have no co-regulating transcription factors at all.

Table 4. Homologous genes regulated by the same transcription factor

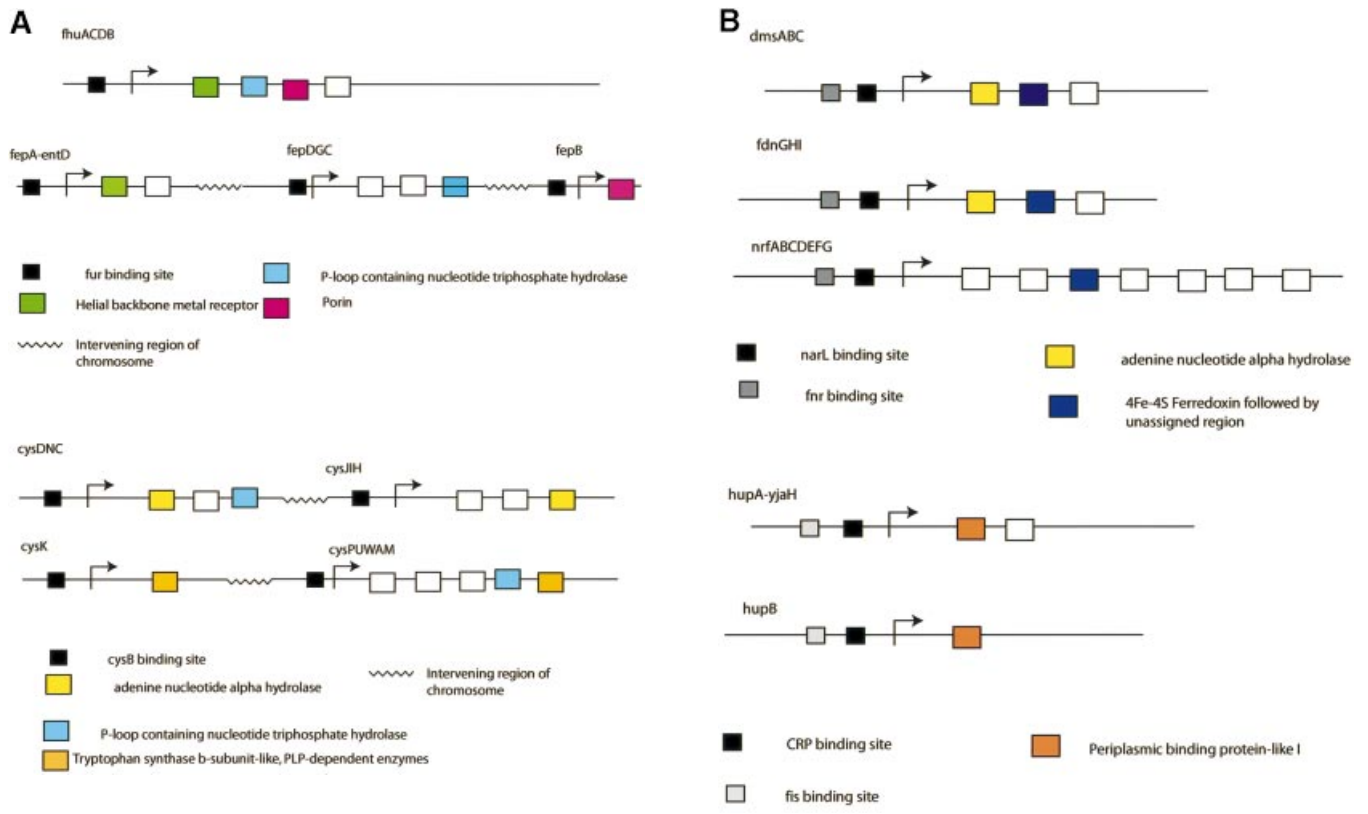
Combination of transcription factors regulating operons	Homologous gene(s) in first operon	Homologous gene(s) in second operon	
BirA	BioF	BioA	
Fnr	ArcA	NarL	
HimA	HimA	HimD	
CysB	CysA	CysC	Figure 6A
CysB	CysM	CysK	Figure 6A
CysB	CysH	CysD	Figure 6A
TyrR	AroF	AroG	
ArgR	ArgF	ArgI	
Fur	FepC	FhuC	Figure 6A
Fur	FepB	FhuD	Figure 6A
Fur	FepA	FhuA	Figure 6A
Hns	MukB	ProV	
Phob	PhnCDKLN	PstBS	
PhoB	PhnCKLN	PhoH	
CpxR	LpxD	LpxA	
LexA	RecN	UvrD	
PurR	PurD	PurK	
Crp, Fis	HupA	HupB	Figure 6B
Fnr, NarL	DmsAB	FdnGH	Figure 6B
Fnr, NarL	DmsB	NrfC	Figure 6B

transcription factors, one by six and four operons known to be controlled by five transcription factors, as shown in Figure 5. An example of a gene regulated by several transcription factors is the transcription factor tdcA in Figure 3, which is controlled by five different proteins apart from itself. tdcA is part of the threonine dehydratase operon, with seven genes which are involved in carbon compound metabolism (primarily growth).

There is evidence from other organisms that the same pair of transcription factors has adjacent binding sites in many regulatory regions in the genome. In yeast, such synergistic pairs of transcription factors were studied by Pilpel *et al*. (33), and Berman *et al*. (34) analysed clustered binding sites of five transcription factors active in the early *Drosophila* embryo. In the current data set, 24 pairs of transcription factors regulate between two and five operons, and four triplets of transcription factors regulate two or three operons.

Table 3 shows the distribution of transcription factors that are present at the same promoter as other transcription factors. Seven transcription factors occur at the same promoter with over 10 different transcription factors, and CRP occurs with 52 other transcription factors. This large variation of combinations suggests that not all, if any, of these transcription factors interact physically in a specific manner, such that the interactions with the DNA and the RNA polymerase are the decisive ones. This is supported by experiments such as those of Martin *et al*. (35).

Genes with similar regulatory regions. Given that regulatory regions are composed of binding sites for one or more transcription factors as described above, we want to address the evolution of the regulatory regions by looking for evidence of duplications of regulatory regions with their regulated genes. We define such duplications as operons of homologous genes which are regulated by the same transcription factor(s). Homologous genes are those whose protein products have the same domain architecture according to SUPERFAMILY domain assignments.



Twenty such cases are shown in Table 4. A rough indication of the duplication rate of operons and their regulatory regions can be obtained by dividing these 20 duplicates by the 303 operons with genes with structural assignments in our set. This is a 7% duplication rate, compared to the three-quarters of transcription factors that have evolved by gene duplication. Since a transcription factor regulates genes that are functionally and not evolutionarily related, one would not expect a particularly high level of duplication of regulatory regions together with downstream operons. However, the results presented here and in Rajewsky *et al*. (36) show that duplication does contribute to the evolution of the regulatory network.

Two examples of individual transcription factors regulating homologous genes are given in Figure 6A. Figure 6B shows cases of pairs of transcription factors regulating homologous genes. In the case of hupA and hupB shown in Figure 6B, both have one binding site for CRP and four for FIS. However, the four FIS-binding sites of hupA are all upstream of the transcription start and FIS is an activator, while for hupB, one of the sites is downstream and FIS is a repressor. In fact, the numbers and positions of transcription factor-binding sites are actually very different in nine of the 20 cases of homologous genes regulated by the same transcription factor(s).

Homologous transcription factor-regulated gene modules. In the previous section, we considered similar regulatory regions and possible duplications of genes together with their regulatory regions. We can extend this to combinations of a transcription factor and regulated gene that are both homologous to another transcription factor and its regulated gene. This would provide evidence for growth of the regulatory network through duplication of sections of the chromosome that include a transcription factor and its regulated gene.

The sets of transcription factors and regulated genes that have homologues are shown in Table 5. In the cases of autoregulation, the transcription factor and regulated gene are the same. There are 28 transcription factors with regulated genes that may have evolved by duplication out of 303 sets in total. This represents a duplication level of 9%, which is very small compared to gene duplication levels amongst the transcription factors or compared to domain duplication levels generally found in genomes (23).

Two examples of module duplication are given in Figure 6C. In both of the examples, the transcription factors are located next to the regulated genes on the chromosome, suggesting that the regulatory module may have duplicated as one unit. In both cases, the arrangement of the operons is slightly different. In addition, the numbers and positions of transcription factor-binding sites is different, though this is not shown in the figure. Based on the data we have here, duplication of both regulatory regions and genes, or of transcription factors together with regulated genes, plays a minor role in the evolution of the gene regulatory network in *E.coli*.

DISCUSSION AND CONCLUSIONS

By using domains assigned to *E.coli* transcription factors through homology to proteins of known three-dimensional structure, we can accurately determine the

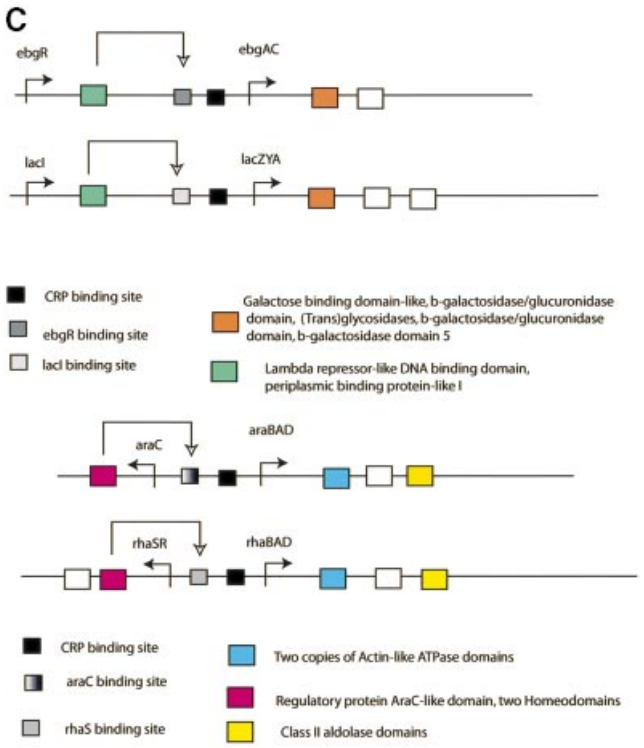


Figure 6. (Opposite and above) Duplication of regulatory modules in the network. (**A**) Two sets of homologous genes regulated by the same transcription factor. In the first example, three genes forming part of the same operon are homologous to three genes in separate operons, all regulated by fur. (**B**) Two examples of homologous genes regulated by the same pair of transcription factors. In the second example, even though hupA and hupB are both regulated by FIS, this transcription factor is an activator in one case and a repressor in another. (**C**) Two examples of duplication of regulatory modules of transcription factors and regulated genes. All genes are involved in the breakdown of various sugars.

Table 5. Homologous transcription factors and regulated genes

Transcription factors	Homologous regulated genes	Transcription factors	Homologous regulated genes
Crp, GalS	MglAB	Crp, RbsR	RbsAB
Crp, GalS	MglA	Crp, GntR	GntK
Crp, EbgR	EbgA	Crp, LacI	LacZ
Crp, RhaS	RhaBD	Crp, AraC	AraBD
Crp, FucR	FucK	Crp, YiaJ	LyxK
Crp, FucR	FucA	Crp, YiaJ	SgbE
RcsB	WcaB	CpxR	LpxA
RcsB	WcaB	CpxR	LpxD
RcsB	B2060	PhoB	PhnCKLN
RcsB	B2060	PhoB	PstB
RcsB	B2060	PhoB	PhoH
MhpR	MhpF	Fur	EntA
FecI, Fur	FecI	UidR, UxuR	UidR
Crp, FucR	FucR	Crp, Fur	Fur
Homologous pairs of transcription factors with autoregulation			
IdnR	PurR		
AsnC	Lrp		
Fnr	Crp		
Homologous triplets of transcription factors with autoregulation			
TorR	PhoB	CpxR	
EmrR	PdhR	ExuR	
Eight homologous transcription factors with autoregulation			
DsdC	GcvA	CynR	Hcar
IlvY	LysR	OxyR	CysB

domain architectures and evolutionary relationships of this repertoire of proteins for the first time. Previous analyses of large numbers of predicted *E.coli* transcription factors have focused on the small helix–turn–helix motif, and largely neglected the fact that this is part of different families of DBDs and their combinations with different families of partner domains (2,9). With our approach of using structural assignments to *E.coli* proteins, we identified 271 proteins as transcription factors with DBDs. This set is likely to represent a large fraction of all transcription factors.

This set of 271 transcription factors identified by us have DBDs from 11 different families. About three-quarters of these transcription factors are two-domain proteins with one DBD and one control domain, which, most frequently, is a small molecule-binding domain. By grouping the transcription factors according to their domain architectures, we found that almost three-quarters of the transcription factors have evolved as a consequence of complete gene duplication.

The rate of duplication of regulatory modules is much lower: only 7% of regulated operons have homologous genes regulated by the same transcription factor and 9% of transcription factor-regulated operon modules have homologues. This suggests that the individual elements of transcription factors, regulatory regions and regulated genes mainly evolve separately.

The set of transcription factors can be classified into 10 broad functional classes, and in certain of these there are global regulators that control many genes. Eight of the 11 global regulators are in the following four categories: carbon degradation, carbon metabolism, redox sensing and control of ion transport. The global regulators amplify their effect by regulating other transcription factors, and overall 38 of the 120 transcription factors with regulatory information control other transcription factors. These transcription factors that regulate other transcription factors are collated in a single figure, including information about the DBD family of the proteins. This figure provides a summary of the central part of the transcriptional network currently known in *E.coli*, and reveals that there are a small number of multi-level regulatory cascades amongst the transcription factors.

Supplementary data

The set of 271 transcription factors and their domain assignments is available at http://www.mrc-lmb.cam.ac.uk/genomes/madanm/ec_tf/.

ACKNOWLEDGEMENTS

We acknowledge Julio Collado-Vides and Heladia Salgado for readily providing us with information from RegulonDB, Julian Gough for help with the SUPERFAMILY structural assignments and Cyrus Chothia, Graeme Mitchison and Andrew Travers for comments on the manuscript. We are grateful to the Medical Research Council, Cambridge Commonwealth Trust and Trinity College, Cambridge, for financial support.

REFERENCES

1. Perez-Rueda,E. and Collado-Vides,J. (2000) The repertoire of DNA-binding transcriptional regulators in *Escherichia coli* K-12. *Nucleic Acids Res.*, **28**, 1838–1847.
2. Perez-Rueda,E. and Collado-Vides,J. (2001) Common history at the origin of the position-function correlation in transcriptional regulators in archaea and bacteria. *J. Mol. Evol.*, **53**, 172–179.
3. Thieffry,D., Huerta,A.M., Perez-Rueda,E. and Collado-Vides,J. (1998) From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli*. *Bioessays*, **20**, 433–440.
4. Shen-Orr,S.S., Milo,R., Mangan,S. and Alon,U. (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genet.*, **31**, 64–68.
5. Guelzim,N., Bottani,S., Bourgine,P. and Kepes,F. (2002) Topological and causal structure of the yeast transcriptional regulatory network. *Nature Genet.*, **31**, 60–63.
6. Morett,E. and Segovia,L. (1993) The sigma 54 bacterial enhancer-binding protein family: mechanism of action and phylogenetic relationship of their functional domains. *J. Bacteriol.*, **175**, 6067–6074.
7. Aravind,L. and Koonin,E.V. (1999) DNA-binding proteins and evolution of transcription regulation in the archaea. *Nucleic Acids Res.*, **27**, 4658–4670.
8. Jacob,F. and Monod,J. (1961) Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3**, 318–356.
9. Rosinski,J.A. and Atchley,W.R. (1999) Molecular evolution of helix-turn-helix proteins. *J. Mol. Evol.*, **49**, 301–309.
10. Gough,J., Karplus,K., Hughey,R. and Chothia,C. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.*, **313**, 903–919.
11. Gough,J. and Chothia,C. (2002) SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res.*, **30**, 268–272.
12. Murzin,A.G., Brenner,S.E., Hubbard,T.J. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
13. Lo Conte,L., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.*, **30**, 264–267.
14. Krogh,A., Brown,M., Mian,I.S., Sjolander,K. and Haussler,D. (1994) Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.*, **235**, 1501–1531.
15. Madera,M. and Gough,J. (2002) A comparison of profile hidden Markov model procedures for remote homology detection. *Nucleic Acids Res.*, **30**, 4321–4328.
16. Karplus,K., Barrett,C. and Hughey,R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.
17. Riley,M. (1998) Genes and proteins of *Escherichia coli* K-12. *Nucleic Acids Res.*, **26**, 54.
18. Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. and Koonin,E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.
19. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.
20. Salgado,H., Santos-Zavaleta,A., Gama-Castro,S., Millan-Zarate,D., Diaz-Peredo,E., Sanchez-Solano,F., Perez-Rueda,E., Bonavides-Martinez,C. and Collado-Vides,J. (2001) RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucleic Acids Res.*, **29**, 72–74.
21. Falconi,M., Brandi,A., La Teana,A., Gualerzi,C.O. and Pon,C.L. (1996) Antagonistic involvement of FIS and H-NS proteins in the transcriptional control of hns expression. *Mol. Microbiol.*, **19**, 965–975.
22. Gonzalez-Gil,G., Kahmann,R. and Muskhelishvili,G. (1998) Regulation of crp transcription by oscillation between distinct nucleoprotein complexes. *EMBO J.*, **17**, 2877–2885.
23. Teichmann,S.A., Park,J. and Chothia,C. (1998) Structural assignments to the *Mycoplasma genitalium* proteins show extensive gene duplications and domain rearrangements. *Proc. Natl Acad. Sci. USA*, **95**, 14658–14663.
24. Qian,J., Luscombe,N.M. and Gerstein,M. (2001) Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *J. Mol. Biol.*, **313**, 673–681.
25. Kwon,H.J., Bennik,M.H., Demple,B. and Ellenberger,T. (2000) Crystal structure of the *Escherichia coli* Rob transcription factor in complex with DNA. *Nature Struct. Biol.*, **7**, 424–430.
26. Rhee,S., Martin,R.G., Rosner,J.L. and Davies,D.R. (1998) A novel DNA-binding motif in MarA: the first structure for an AraC family transcriptional activator. *Proc. Natl Acad. Sci. USA*, **95**, 10413–10418.
27. Anantharaman,V., Koonin,E.V. and Aravind,L. (2001) Regulatory potential, phyletic distribution and evolution of ancient, intracellular small-molecule-binding domains. *J. Mol. Biol.*, **307**, 1271–1292.
28. Apic,G., Gough,J. and Teichmann,S.A. (2001) Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol.*, **310**, 311–325.
29. Bashton,M. and Chothia,C. (2002) The geometry of domain combination in proteins. *J. Mol. Biol.*, **315**, 927–939.
30. Madan Babu,M. and Teichmann,S.A. (2003) Functional determinants of transcription factors in *Escherichia coli*: protein families and binding sites. *Trends Genet.*, **19**, 75–79.
31. Rosenfeld,N., Elowitz,M.B. and Alon,U. (2002) Negative autoregulation speeds the response times of transcription networks. *J. Mol. Biol.*, **323**, 785–793.
32. Brandi,A., Pon,C.L. and Gualerzi,C.O. (1994) Interaction of the main cold shock protein CS7.4 (CspA) of *Escherichia coli* with the promoter region of hns. *Biochimie*, **76**, 1090–1098.
33. Pilpel,Y., Sudarsanam,P. and Church,G.M. (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genet.*, **29**, 153–159.
34. Berman,B.P., Nibu,Y., Pfeiffer,B.D., Tomancak,P., Celniker,S.E., Levine,M., Rubin,G.M. and Eisen,M.B. (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl Acad. Sci. USA*, **99**, 757–762.
35. Martin,R.G., Gillette,W.K., Martin,N.I. and Rosner,J.L. (2002) Complex formation between activator and RNA polymerase as the basis for transcriptional activation by MarA and SoxS in *Escherichia coli*. *Mol. Microbiol.*, **43**, 355–370.
36. Rajewsky,N., Socci,N.D., Zapotocky,M. and Siggia,E.D. (2002) The evolution of DNA regulatory regions for proteo-gamma bacteria by interspecies comparisons. *Genome Res.*, **12**, 298–308.