Research

Conservation of the binding site for the arginine repressor in all bacterial lineages

Kira S Makarova*†, Andrey A Mironov‡ and Mikhail S Gelfand ‡

Addresses: *Department of Pathology, Uniformed Services University of the Health Sciences, Bethesda, MD 20814, USA and National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA. ‡State Scientific Center GosNIIGenetika, Moscow 113545, Russia.

†Permanent address: Institute of Cytology and Genetics, Russian Academy of Sciences, Novosibirsk 630090, Russia.

Correspondence: Mikhail S Gelfand. E-mail: misha@imb.imb.ac.ru

Published: 22 March 2001

Genome ***Biology*** 2001, **2(4)**:research0013.1–0013.8

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/2/4/research/0013

Received: 30 October 2000
Revised: 14 December 2000
Accepted: 6 February 2001

(Print ISSN 1465-6906; Online ISSN 1465-6914)

Abstract

Background: The arginine repressor ArgR/AhrC is a transcription factor universally conserved in bacterial genomes. Its recognition signal (the ARG box), a weak palindrome, is also conserved between genomes, despite a very low degree of similarity between individual sites within a genome. Thus, the arginine repressor is different from two other universal transcription factors - HrcA, whose recognition signal is very strongly conserved both within and between genomes, and LexA/DinR, whose signal is strongly conserved within, but not between, genomes. The arginine regulon is well studied in *Escherichia coli* and to some extent in *Bacillus subtilis* and some other genomes. Here, we apply the comparative genomic approach to the prediction of the ArgR-binding sites in all completely sequenced bacterial genomes.

Results: Orthologs of ArgR/AhrC were identified in the complete genomes of *E. coli*, *Haemophilus influenzae*, *Vibrio cholerae*, *B. subtilis*, *Mycobacterium tuberculosis*, *Thermotoga maritima, Chlamydia pneumoniae* and *Deinococcus radiodurans*. Candidate arginine repressor binding sites were identified upstream of arginine transport and metabolism genes.

Conclusions: We found that the ArgR/AhrC recognition signal is conserved in all genomes that contain genes encoding orthologous transcription factors of this family. All genomes studied except *M. tuberculosis* contain ABC transport cassettes (related to the Art system of *E. coli*) belonging to the candidate arginine regulons.

Background

Bacterial and archaeal transcriptional regulators typically form large protein families consisting of numerous paralogs (for example the LacI/GntR, AraC and DeoR families [1]). Only three readily detectable clusters of orthologous transcription factors include just one or two representatives from a broad range of diverse branches of bacteria, namely the SOS repressors LexA/DinR, the heat-shock repressor HrcA, and the arginine repressor ArgR/AhrC [2] (Table 1). A comparison of the coevolution of these conserved regulators and their binding sites in DNA could reveal general trends in the evolution of regulons.

Table 1

Comparison of three transcriptional regulator families with predominantly single representatives from each bacterial genome

Definition	Regulated pathway	Pattern of species*	Type of DNA-binding domain and fused domain	DNA-binding domain conservation†	Recognition site‡	Sites per genome	Reference§
LexA (Gram-negative) DinR (Gram-positive)	SOS repair	ADC-VEBMH----	'Winged helix'¶ HTH and serine protease (S24 family)	2.20 ± 0.28	SOS box (Gram-negative): CTGTatatatatMCAG Cheo box (Gram-positive): cGAACrnryGTTYg	8-20	[3,4]
HrcA	Heat shock	A-CS-EBM-XYTP	Predicted 'winged helix' HTH and uncharacterized domain possibly responsible for activation by chaperonin GroE	1.72 ± 0.22	CIRCE box: TTAGCACTCn_9GAGTGCTAA	1-2	[5,6]
ArgR (Gram-negative) AhrC (Gram-positive)	Arginine metabolism	ADC-VEBMH---P	'Winged helix' HTH and arginine-binding domain	1.81 ± 0.22	ARG box: TGMATwwwwATKCA	1-20	[8-11]

*Abbreviations for species: B, *Bacillus subtilis*; C, *Clostridium acetobutylicum*; M, *Mycobacterium tuberculosis*; D, *Deinococcus radiodurans*; A, *Thermotoga maritima*; S, *Synechocystis* sp; H, *Haemophilus influenzae*; E, *Escherichia coli*; P, *Chlamydia pneumoniae*; T, *Chlamydia trachomatis*; Z, *Mycoplasma genitalium*; Y, *Mycoplasma pneumoniae*; V, *Vibrio cholerae*. †The estimate was obtained as the average maximum likelihood distance between the last-step UPGMA clusters (of the corresponding tree reconstructed by the PHYLIP package program NEIGHBOR) counted using distance matrix (calculated by PHYLIP package program PROTDIST) only for DNA-binding domains [39]. ‡ Letter codes used in consensus sequences are the following: M = A or C; Y = T or C; R/r =A or G; W/w =A or T; K = G or T; N/n = any nucleotide, where upper-case letters denote strongly conserved nucleotides and lower-case letters denote less conserved nucleotides. § References correspond to the two last columns. ¶The 'winged helix' superfamily is defined in the SCOP database [40]; LexA [41] and ArgR [42-44] DNA-binding domains have been resolved by X-ray crystallography. The same type of domain was predicted for HrsA using PSIBLAST program.

The signals recognized by LexA in Gram-negative bacteria and by its ortholog DinR in Gram-positive bacteria (the SOS box [3] and the Cheo box [4], respectively) are completely different. Accordingly, the DNA-binding domains of these proteins are divergent (Table 1). The heat-shock regulator HrcA binds CIRCE elements that are located upstream of genes encoding heat-shock proteins (molecular chaperones) in many different genomes [5,6]; in the mycoplasmas, HrcA also regulates heat-shock protease genes [7]. The CIRCE signal is very specific (two complementary nonamers with a 9 base pair (bp) spacer) and is extremely highly conserved in all genomes that encode HrcA (not more than five, and usually less than three, mismatches to the consensus in all known and predicted sites [7]). The amino acid sequence of HrcA is conserved as well (Table 1).

The arginine regulon, which is regulated by the arginine repressor ArgR/AhrC, represents an evolutionary strategy distinct from that of either the SOS or the heat-shock regulons. The DNA-binding domains of the ArgR/AhrC family are less conserved than those of the HrcA family, but more conserved than those of the LexA/DinR family (Table 1, column 5). DNA signals recognized by ArgR/AhrC are also similar in several bacterial lineages at least [8-11]. These sites often occur in pairs [12-15], although single-box sites have also been shown to bind ArgR/AhrC, for example the sites in the catabolic operons of *B. subtilis* [9], the adenine deaminase pathway operon in *Bacillus licheniformis* [14], and the *cer* recombination region of the *E. coli* plasmid ColE1 ([16,17]; see also the study of mutated ArgR [18]). Unlike the CIRCE element, the ARG box seems to be weakly conserved, even within a genome, and the specificity of recognition is often achieved by cooperative interactions between tandem sites, as shown in both experimental [9,12,13] and statistical [19] studies. The set of ARG boxes from different genomes, however, is fairly homogeneous, and indeed, arginine repressors from different bacteria appear to be at least partially interchangeable within major taxonomic groups: there is some cross-binding between ArgR and AhrC [20]; ArgR but not AhrC binds to the *Thermus thermophilus* sites [21] and AhrC binds to the *Streptomyces coelicolor* sites [22]. The ARG box consensus was described as TNTGAATWWWWATTCANW in *E. coli* [8,12], CATGAATAAAAATKCAAK in *B. subtilis* [9,10] and AWTGCATRWWYATGCAWT in Streptomycetes [11] (where W = A or T, K = G or T, R = A or G, Y = T or C, N = any base; Table 1). In addition, binding of ArgR homologs to the sites similar to ARG boxes was reported for other *Bacillus* species (*B. licheniformis* [14] and *B. stearothermophilus* [23,24]), and for *Salmonella typhimurium* [25]. Several ArgR-binding sites were predicted on the basis of similarity with the *E. coli* consensus in the upstream regions of various genes involved in arginine metabolism in *Moritella* [26].

In a previous study [27], we used comparative genomic analysis of regulatory signals to predict the gene composition of the arginine regulon of *Haemophilus influenzae* using the well characterized *E. coli* regulon as the starting

Table 2

Candidate ARG boxes upstream of arginine metabolism related genes and operons

Genome	Operon	Position	Score	Site
E. coli	*argR*	-64	4.24	**ttTGcATAAAAATTCATc**
		-43	3.34	**tATGcAcAAtAATgttgT**
	argA	-50	3.98	**AcaGAATAAAAATaCAcT**
		-39	3.98	**ttcGAATAAtcATgCAaa**
	argCBH	-128	4.61	**tATcAATAttcATgCAgT**
		-109	4.61	**tATGAATAAAAATaCAcT**
	argD	-68	4.01	**AgTGAtTttttATgCATa**
		-47	3.50	**tgTGgtTAtAAtTTCAca**
	argE	-64	3.80	**AgTGtATttttATTCATa**
		-43	3.39	**AcTGcATgAAtATTgATa**
	argF	-65	4.16	**AATGAATAAttAcaCATa**
		-44	4.41	**AgTGAATtttAATTCAaT**
	argG	-210	4.31	**tgTGAATgAAtATcCAgt**
		-91	3.90	**AtTaAATgAAAAcTCATT**
		-70	4.51	**ttTGcATAAAAATTCAgT**
	argI	-63	4.33	**AATGAATAAtcATcCATa**
		-42	4.49	**AtTGAATtttAATTCATT**
	carAB	-50	4.36	**tgTGAATtAAtATgCAaa**
		-39	3.79	**AgTGAgTgAAtaTTCtcT**
	artPIQM	-72	4.08	AtTGaATAAttATTCtgT
	artJ	-86	4.36	AtTGcATAtAAATTCAcT
H. influenzae	*HI1209*	-50	4.27	AgTGAATttttATgCAaT
	HI0811	-54	4.52	tATGAATAAAtATgCAca
	HI1727	-64	3.87	AtaGAATttttATTCAca
		-43	3.75	AtcGAtTAtttATTCAaT
	HI1180-77	-50	4.01	tATGcATAAAAATgtAaT
V. cholerae	*VC2316*	-15	3.82	AaaGAATAAAAAgTCATT
	VC2390-89	-52	3.57	ttTGcAaAAtAATTtATT
	VC2508	-72	3.72	ttaaAATAtttATTCAcT
		-51	3.35	AtaGcATtttcATgCtTT
	VC2618	-120	3.83	AaTaAATgtAAATaCAaT
	VC2644-42	-76	3.00	AacacATAttAAaTCAcT
		-55	4.00	AGTGAATAAAAAaaCAaT
	VC2645	-79	3.50	AtTGttTttttATTCAcT
		-58	3.27	AGTGAtTtAAtATgtgTT
	VCA0757-60	-74	3.00	ttTtggTttttATaCATT
		-53	3.85	AtTGcATAAAAATaCgTT
B. subtilis	*argCJBD*	-64	5.22	**ATtAATTTtTATTCAT**
	carABargF	-55	4.54	AgGcATAaAAATTCAT
		-35	4.02	AtaAtTAatTATTCAT
	argGH	-67	5.04	**ATGtATTTtTATTaAa**
	yqjN	-31	4.54	AaGcATTTtTATTCAT
		-47	4.91	ATttATTTtTATaCAa
	rocABC	-53	5.14	**tTGcATTTtTATTCAT**
	rocDEF	-63	4.76	**tTGcATTTATATaaAg**
	rocR	-193	4.76	cTttATATAAATgCAa
	yqiXYZ	-88	4.21	tTGcATAaAAATgaga
		-51	4.62	AcGAATAaATATTCAa
C. acetobutylicum	*414*	-191	5.27	ATGAATAaATATTCAa
	491	-132	4.62	tTGAATAaATATTCgT
		-32	5.27	ATGAATAaAAATTCAa
	1203-4	-124	5.04	ATGAATATtTATaaAa
	2787-8	-63	4.79	tTGAATATtTATaaAg
		-43	4.59	gTGtATAatTATTCAT
	2786-2785	-118	4.59	ATGAATAatTATaCAc
		-98	4.79	cTttATAaATATTCAa
	3090-89	-181	4.96	ATGcATAaATATaaAT
		-80	4.91	ATGtATAaATATaaAa
	3533	-87	4.96	ATGcATAaATATaaAT
		-67	4.89	tTGAATAatTATaaAa
	4268-65	-194	5.09	ATGcATAaATATTaAT
		-69	4.62	gTtAATAatTATTCAT

Table 2 (continued)

Genome	Operon	Position	Score	Site
T. maritima	*TM0371*	-59	4.21	tTtcATATtTATgCta
	TM0558-57	-81	3.92	tTtAATTcAAAgTaAa
	TM0593-91	-34	4.16	tTGtgTTatAATaaAT
	TM1780-85	-138	3.87	cgtAATTgATATTCAT
	TM1873	-91	4.44	ATttATTTAAcTTaAT
D. radiodurans	*DR0080*	-57	4.11	cTGtATTTcTATaCAg
	DR0674-78	-131	4.09	tTGcATAgtcATTCAT
	DR2610	-85	3.62	ATGgATTgAAATcCAg
		-62	3.69	cTGgATTTtAAggaAT
C. pneumoniae	*glnPQ*	-53	4.26	tTGcATAaATATgatT
		-32	4.64	ATaAATAaATATgCAT
	artJ	-112	3.92	tTtAATcaAAATTatT
		27	3.96	ATttATTTtTATaatg
M. tuberculosis	*Rv1652-59*	-11	3.91	tTGcATAacgATgCAa

Position is indicated relative to the start of translation. The score for *E. coli* and *H. influenzae* genes was computed using the profile from [27]; other scores were computed using the profile trained on *B. subtilis* candidate ARG boxes using the procedure from [28]. The sites used to construct the profile are shown in bold.

point. Here we extend this analysis to explore the conservation of the ARG box in all bacteria that encode an ortholog of the ArgR repressor.

Results and discussion

The comparative approach to the analysis of regulation is based on the assumption that regulons (sets of co-regulated genes) are conserved in genomes containing orthologs of the relevant regulatory proteins. Thus true candidate binding sites for the regulator occur upstream of orthologous genes, whereas false positives are scattered at random in the genome. This provides a consistency check that sharply increases the accuracy of prediction.

The ARG box profile constructed as described in the Metraisl and methods section was used to scan the complete genomes of other bacteria (excluding the gamma-proteobacteria). The profile is not very selective: at threshold z-score = 3.75 [27] about 1% of the *B. subtilis* and *M. tuberculosis* genes are selected, compared with 7% for *T. maritima*. Nevertheless, there is a sharp distinction between the arginine-related genes without ARG boxes (for example, *argT* of *E. coli*, *argF* of *H. influenzae*, *carAB* of *M. tuberculosis*, *argF* of *T. maritima* and several *Deinococcus* genes, see Figure 1) and those with relatively strong and probably functional ARG boxes. Only the genes involved in arginine metabolism and transport (see below) have upstream ARG boxes in more than five out of eight of the genomes considered. Thus despite the seeming weakness of individual predictions, the basic assumption of the regulon conservation yields validity of the candidate sites [27,28]. Many weaker sites are second sites in cooperative cassettes. The candidate ArgR-binding sites are listed in Table 2 and shown in Figure 1. Validity of the

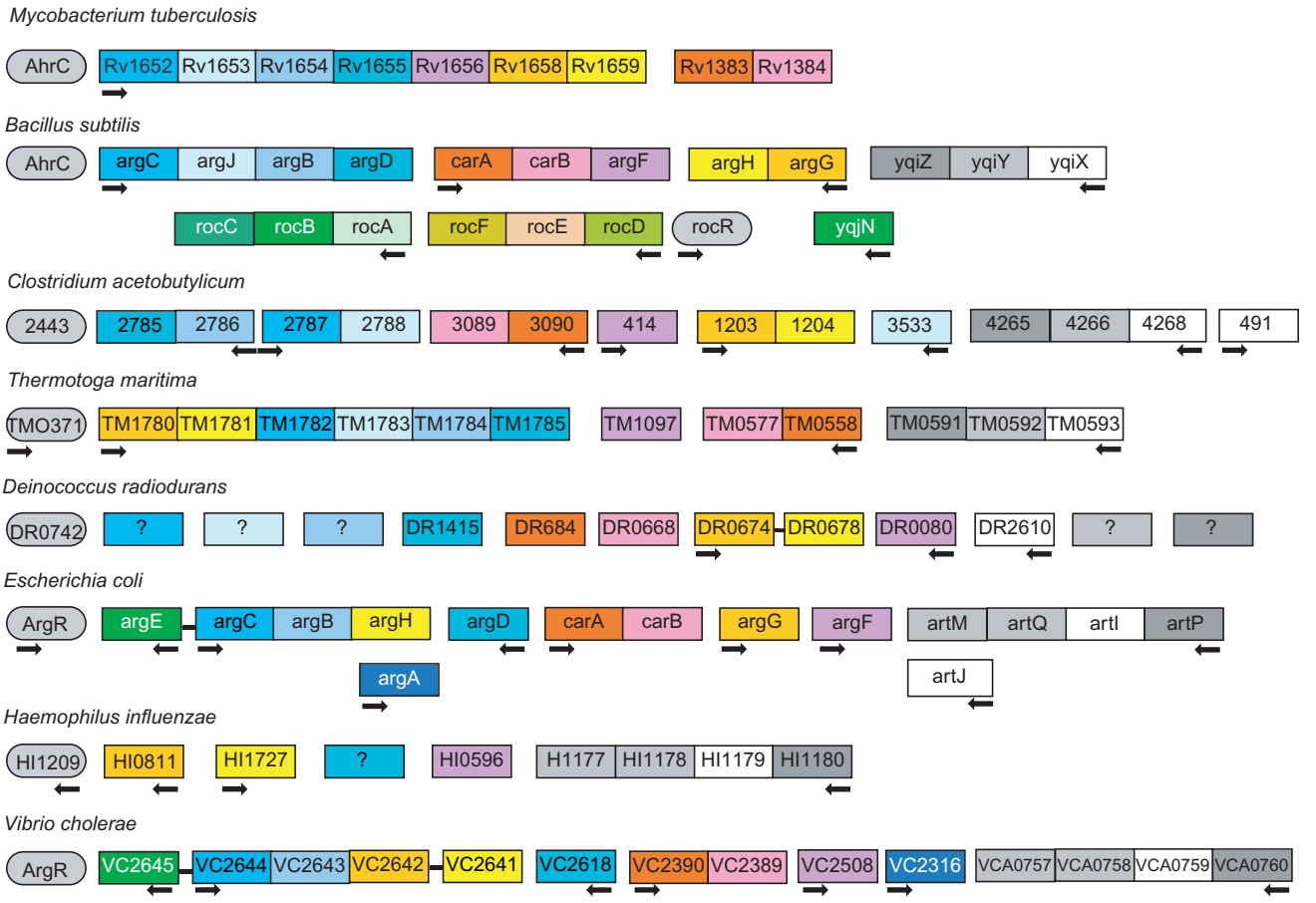

Figure 1
Schematic representation of the operon organization and regulation of the arginine metabolism and transport genes. Genes are represented by boxes. ARG boxes in the upstream region are shown by black arrows. The direction of the arrow indicates the direction of transcription. The linear pathway (in *E. coli* and *V. cholerae*) involves *N*-acetylglutamate synthase (*argA*) and *N*2-acetylornithine deacetylase (*argE*). The circular pathway (in other bacteria) involves *N*2-acetyl-L-ornithine: L-glutamate acetyltransferase (*argJ*). The common genes are acetylglutamate kinase (*argB*); acetylglutamate semialdehyde dehydrogenase (*argC*); acetylornitine delta-aminotransferase (*argD*); ornithine carbamoyltransferase (*argF*, *argI*); argininosuccinate synthase (*argG*); argininosuccinate lyase (*argH*); carbamoyl-phosphate synthase (*carAB*). The *H. influenzae* genome contains only *argH*, *argG*, *argF* and possibly *argD* orthologs. There are difficulties in identifying orthologs for *argC*, *argJ* and *argB* in *D. radiodurans* because there are several paralogous genes encoding proteins that can possibly perform these functions. The *B. subtilis roc* operons involved in arginine degradation are also regulated by AhrC, as well as anaerobic arginine catabolism genes *arcABCD* in *B. licheniformis* [14] (data not shown). The transporter genes are: periplasmic binding protein (white), permease transmembrane protein (light gray), ATPase component (dark gray).

B. subtilis profile for analysis of other genomes is confirmed by a candidate ARG box with z-score = 3.96 within the region protected when ArgR binds upstream of the *argR* gene of *Thermotoga neapolitana* [29] (data not shown).

In addition to previously characterized ARG boxes in *B. subtilis* we identified a candidate ARG box upstream of the *yqjN* gene (Figure 1, Table 2), a probable product of recent duplication of the *rocB* gene encoding an arginine utilization protein with unknown biochemical function. Thus is it likely that YqjN has the same function as RocB and is also involved in arginine degradation.

An important outcome of the analysis is that in addition to the genes encoding the arginine metabolism enzymes, ArgR probably regulates ABC-cassette operons or scattered genes responsible for arginine transport in all bacteria except *M. tuberculosis* and maybe *C. pneumoniae* (Figure 1). Straightforward resolution of the orthology relationships between genes involved in transport of polar amino acids on the basis of their sequence similarity is impossible (Figure 2, and see COG0834, COG0795, COG1126 in [1]). Therefore the presence of candidate ARG boxes upstream of these genes could be the only indication of their involvement in arginine transport before experimental verification. Nevertheless, the

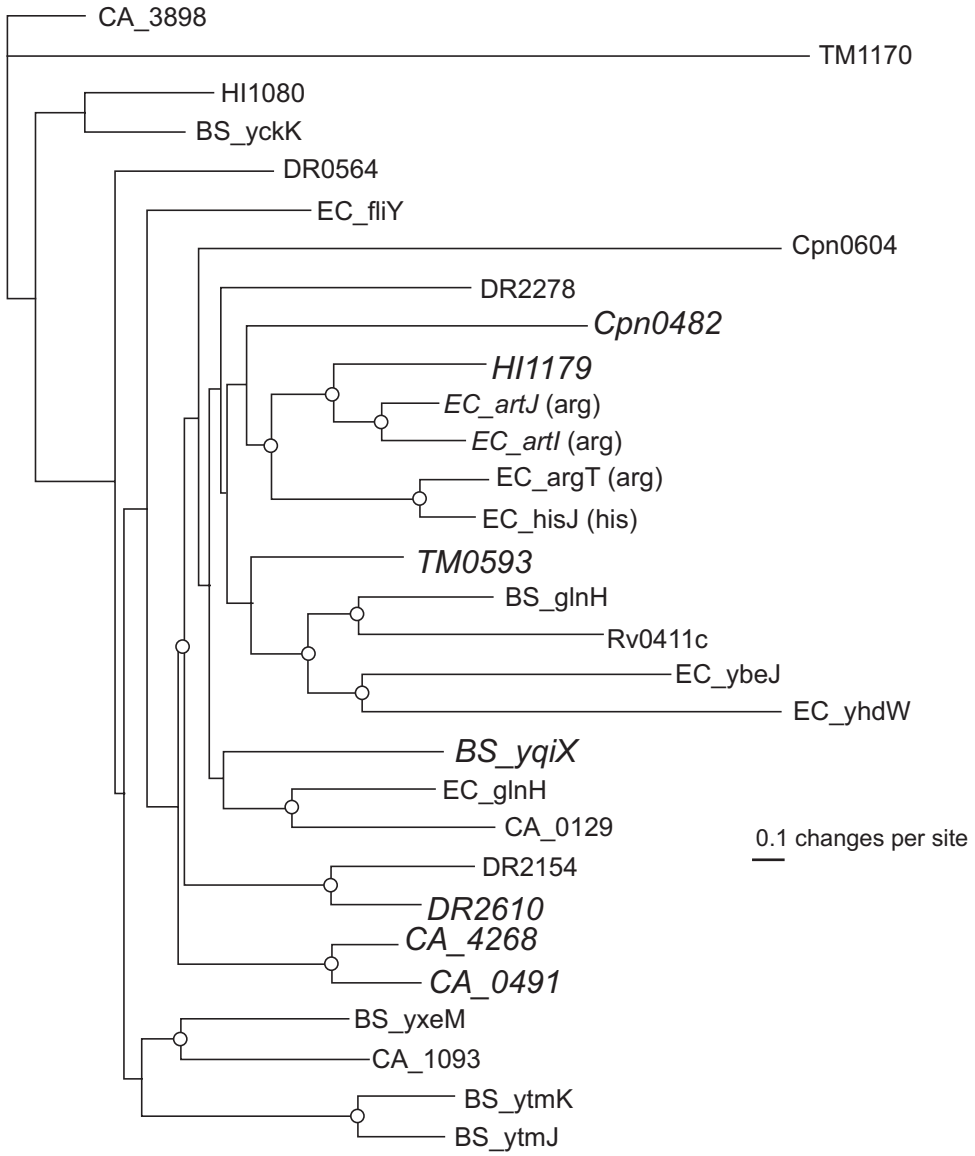

Figure 2
Unrooted, neighbor-joining tree of the predicted polar amino acid periplasmic binding proteins for selected organisms. The tree was reconstructed using the PHYLIP package (SEQBOOT, PROTDIST, NEIGHBOR, CONSENSE and FITCH programs). Nodes with bootstrap value exceeding 60% are marked by open circles. BS, *B. subtilis*; CA, *Cl. acetobutylicum*; Cpn, *C. pneumoniae*; DR, *D. radiodurans*; EC, *E. coli*; HI, *H. influenzae*; Rv, *M. tuberculosis*; TM, *T. maritima*. Experimentally established specificity of transporters is indicated in parentheses. Genes with candidate ARG boxes in upstream regions are shown in italic and in a larger font.

protein tree presented in Figure 2 demonstrates clustering of closely related paralogs within one organism (*E. coli*, *Clostridium acetobutylicum*) or orthologs in closely related organisms (*E. coli* and *H. influenzae*) that have upstream candidate ARG boxes (Figure 1, Table 2). In the *E. coli* genome, this family includes two loci, *artPIQM-artJ* and *argT-hisJQMP*. In each case the four-gene operon encodes a complete ABC cassette with two transmembrane components, whereas the single-gene operon encodes an additional periplasmic protein. The *art* genes encode an arginine transport system. The *hisJQMP* operon encodes a histidine-specific ABC cassette, whereas the product of the upstream gene *argT*, lysine-arginine-ornithine-binding periplasmic protein ArgT, can substitute the periplasmic protein HisJ in binding to the membrane component HisP, thus changing the initial histidine transporter specificity [30]. The operons *hisJQMP* and *argT* have no candidate ARG boxes and do not seem to belong to the arginine regulon.

In the *Pseudomonas aeruginosa* genome there are three systems closely related to the above transporters. One is orthologous to *hisJQMP* and the other to *artPIQM*. These two systems have not been characterized experimentally. The third system, *aotQJMP*, is closer to *hisJQMP* than to *artPIQM*. It encodes transporters of arginine and ornithine, but not lysine [31], and is located within the arginine and ornithine catabolism locus *aot-aru*. The *aot* system is positively regulated by an activator, ArgR, which is encoded by the distal gene of the *aotJQMOPargR* operon [31]. This activator belongs to the AraC family and is not related to the ArgR repressor of *E. coli* [32].

The situation with the *C. pneumoniae* genome is not clear. It contains the *argR* gene but no genes for the arginine metabolism. There is a stand-alone *artJ* gene (encoding an ABC-cassette periplasmic protein) and two genes annotated as *glnPQ* immediately downstream of *argR* (encoding the transmembrane and ATPase components respectively). In fact, *glnP* of *C. pneumoniae* is the bidirectional best hit of the *E. coli* gene *yecC* situated in the flagellar locus. The ABC transporters are not easily amenable to orthology analysis, as their specificity may change at a fast rate. As mentioned above, positional and regulatory analysis is often the only computational technique for determining the cellular role of ABC cassettes before experimental verification. We note a pair of ARG boxes upstream of *glnPQ* and two ARG boxes with lower z-scores upstream of the *artJ* operon of *C. pneumoniae*. Thus it is very tempting to predict that these genes in fact encode an arginine transport system regulated by ArgR. We feel, however, that this prediction cannot be accepted without experimental verification, especially in view of two complicating observations. First, both *artJ* and *glnPQ* operons are conserved in the genome of *C. trachomatis*, despite the fact that the latter has no gene for ArgR. Second, ArgR of *C. pneumoniae* is closer to the ArgR of gamma-proteobacteria than to the AhrC/ArgR of Gram-positive bacteria, but nevertheless the ARG boxes of *C. pneumoniae* are visible with the *Bacillus* profile, but not with the gamma-proteobacteria profile.

Taken together these data suggest that ARG regulons represent an interesting (and possibly unique) case which could be considered as an intermediate evolutionary state compared to the HrcA and LexA/DinR regulons. ArgR orthologs retain high similarity on the amino acid level within the major taxonomic groups, and are identifiable between these groups, whereas ARG box conservation is low, although sufficient to be detected in diverged bacterial lineages. Nevertheless, this state seems to be stable and it is not clear what evolutionary forces are responsible for its stability. In this respect it is noteworthy that the structural type of the DNA-binding domain in the protein apparently does not determine the evolutionary relationships with its recognition site. All three aforementioned regulator families, as well as many others, contain the so-called 'winged helix' DNA-binding domain and its conservation is not correlated with conservation of its binding site (Table 1).

Conclusions

The composition of the ARG regulons in different bacteria is known to vary mainly because of diversity in the arginine degradation pathways and species-specific paralogs. The question of the origin of 'additional' ARG boxes thus arises. Because of the low conservation of the ArgR-binding signal, it is possible that some of the sites could be convergent in origin. Moreover, each genome contains a large number of potential ARG box-like sequences that could become actual sites when they become located upstream of an arginine metabolism gene following chromosomal rearrangements [33].

In contrast, CIRCE elements appear to be direct descendants of the ancient regulon present in the common ancestor of the Bacteria, because the variation in the composition of the CIRCE regulon is minimal and the few additional sites found in some genomes are apparently products of duplication. Most other DNA-binding domains of transcriptional regulators (including LexA) seem to undergo considerable changes together with their DNA signals and regulons. Thus, the evolution of the arginine regulon and ARG boxes seems to reflect a tradeoff between maintaining regulon flexibility on one hand and retaining the universal regulatory mechanism on the other.

Another interesting aspect of the arginine regulon strategy is the use of single and cooperative sites. In *E. coli*, the use of cooperative binding sites by ArgR seems to be a consequence of a requirement for a sharper response to a stimulus (arginine starvation) compared to the SOS response (single sites are usually used by LexA) [19]. Unfortunately, the available data seems to be insufficient to draw any systematic conclusions. In particular, as second sites in the cooperative cassettes are often weak (have low scores), some of them could be missed by the recognition rule. Direct experimental studies are needed to clarify this issue. Another problem that was not directly addressed in this study is the role of the *E. coli* arginine repressor in recombination and its binding to the *cer* site, which contains a single ARG box [16,17]. We have noted, however, conservation of this box in the monomerization site *ckr* of the plasmid ColK [34].

There are a few more transcription factor families (biotin operon repressor, COG1654; putative stress-responsive transcriptional regulator PcpC, COG1983; Bvg accessory factor homologs, COG1521 [1]) with a single representative per genome, and it would be interesting to compare them as well. They do not, however, contain a sufficient number of experimentally determined binding sites and are not so ubiquitous in the bacterial genomes as the three regulators discussed previously. With more available genomes, we hope that our approach, combined with positional analysis aimed

at finding co-localized, and thus possibly functionally related enzymes and regulator genes [35,36], will enable us to make this comparison. On the other hand, we feel that the predictions made in this study, especially identification of the Art family ABC transporters in several diverse genomes, are sufficiently interesting to warrant experimental verification.

Materials and methods

The profile for ARG box identification was constructed as follows. Upstream regions of *B. subtilis* operons involved in arginine metabolism were selected. An iterative signal search procedure was applied as described previously [28]. The resulting ARG box profile was constructed using the four sites upstream of *argC*, *argG*, *rocA* and *rocD*. These formally identified sites are a subset of the experimentally known sites [9]. Gamma-proteobacteria were analyzed using the longer *E. coli* ARG box profile taken from [18]. Only genes having candidate sites in five or more out of the eight genomes analyzed were considered as candidate regulon members and were retained for further analysis. This procedure could lead to the loss of some true sites, but ensured that false sites were not accepted.

The complete genomes of *E. coli*, *H. influenzae*, *Vibrio cholerae*, *B. subtilis*, *Mycobacterium tuberculosis*, *Thermotoga maritima*, *Chlamydia pneumoniae* and *Deinococcus radiodurans* were downloaded from GenBank [37]. The complete genome of *Clostridium acetobutylicum* was obtained at [38].

Acknowledgements

We thank Eugene Koonin, Yury Kozlov and Igor Rogosin for useful discussions. This study was partially supported by grants from the Merck Genome Research Institute (244), the Russian Fund of Basic Research (99-04-48247 and 00-15-99362), the Russian State Scientific Program 'Human Genome', INTAS (99-1476), the Howard Hughes Medical Institute (55000309), and Microbial Genome Program, Office of Biological and Environmental Research, DOE (DE-FG02-98ER62583).

References

1. **Phylogenetic classification of proteins encoded in complete genomes** [http://www.ncbi.nlm.nih.gov/COG/]
2. Tatusov RL, Galperin MY, Natale DA, Koonin EV: **The COG database: a tool for genome-scale analysis of protein functions and evolution.** *Nucleic Acids Res* 2000, **28:**33-36.
3. Walker GC: **The SOS response of *Escherichia coli*.** In Escherichia coli *and* Salmonella *in Cellular and Molecular Biology*, Vol 1. Edited by Neidhardt, FC. Washington DC: ASM Press; 1996: 1400-1412.
4. Winterling KW, Chafin D, Hayes JJ, Sun J, Levine AS, Yasbin RE, Woodgate R: **The *Bacillus subtilis* DinR binding site: redefinition of the consensus sequence.** *J Bacteriol* 1998, **180:**2201-2211.
5. Hecker M, Schumann W, Volker U: **Heat-shock and general stress response in *Bacillus subtilis*.** *Mol Microbiol* 1996, **19:**417-428.
6. Segal R, Ron EZ: **Regulation and organization of the *groE* and *dnaK* operons in Eubacteria.** *FEMS Microbiol Lett* 1996, **138:**1-10.
7. Gelfand MS: **Recognition of regulatory sites by genomic comparison.** *Res Microbiol* 1999, **150:**755-771.
8. Maas WK: **The arginine repressor of *Escherichia coli*.** *Microbiol Rev* 1994, **58:**631-640.
9. Miller CM, Baumberg S, Stockley PG: **Operator interactions by the *Bacillus subtilis* arginine repressor/activator, AhrC: novel positioning and DNA-mediated assembly of a transcriptional activator at catabolic sites.** *Mol Microbiol* 1997, **26:**37-48.
10. Klingel U, Miller CM, North AK, Stockley PG, Baumberg S: **A binding site for activation by the *Bacillus subtilis* AhrC protein, a repressor/activator of arginine metabolism.** *Mol Gen Genet* 1995, **248:**329-340.
11. Rodriguez-Garcia A, Ludovice M, Martin JF, Liras P: **Arginine boxes and the *argR* gene in *Streptomyces clavuligerus*: evidence for a clear regulation of the arginine pathway.** *Mol Microbiol* 1997, **25:**219-228.
12. Charlier D, Roovers M, Van Vliet F, Boyen A, Cunin R, Nakamura Y, Glansdorff N, Pierard A: **Arginine regulon of *Escherichia coli* K-12. A study of repressor-operator interactions and of *in vitro* binding affinities versus *in vivo* repression.** *J Mol Biol* 1992, **226:**367-386.
13. Tian G, Lim D, Carey J, Maas WK: **Binding of the arginine repressor of *Escherichia coli* K12 to its operator sites.** *J Mol Biol* 1992, **226:**387-397.
14. Maghnouj A, de Sousa Cabral TF, Stalon V, Vander Wauven C: **The *arcABDC* gene cluster, encoding the arginine deiminase pathway of *Bacillus licheniformis*, and its activation by the arginine repressor *argR*.** *J Bacteriol* 1998, **180:**6468-6475.
15. Wang H, Glansdorff N, Charlier D: **The arginine repressor of *Escherichia coli* K-12 makes direct contacts to minor and major groove determinants of the operators.** *J Mol Biol* 1998, **277:**805-824.
16. Stirling CJ, Szatmari G, Stewart G, Smith MC, Sherratt DJ: **The arginine repressor is essential for plasmid-stabilizing site-specific recombination at the ColE1 cer locus.** *EMBO J* 1988, **7:**4389-4395.
17. Guhathakurta A, Summers D: **Involvement of ArgR and PepA in the pairing of ColE1 dimer resolution sites.** *Microbiology* 1995, **141:**1163-1171.
18. Chen SH, Merican AF, Sherratt DJ: **DNA binding of *Escherichia coli* arginine repressor mutants altered in oligomeric state.** *Mol Microbiol* 1997, **24:**1143-1156.
19. Berg OG: **Selection of DNA binding sites by regulatory proteins: the LexA protein and the arginine repressor use different strategies for functional specificity.** *Nucleic Acids Res* 1988, **16:**5089-6105.
20. Smith MC, Czaplewski L, North AK, Baumberg S, Stockley PG: **Sequences required for regulation of arginine biosynthesis promoters are conserved between *Bacillus subtilis* and *Escherichia coli*.** *Mol Microbiol* 1989, **3:**23-38.
21. Sanchez R, Roovers M, Glansdorff N: **Organization and expression of a *Thermus thermophilus* arginine cluster: presence of unidentified open reading frames and absence of a Shine-Dalgarno sequence.** *J Bacteriol* 2000, **182:**5911-5915.
22. Soutar A, Baumberg S: **Implication of a repression system, homologous to those of other bacteria, in the control of arginine biosynthesis genes in *Streptomyces coelicolor*.** *Mol Gen Genet* 1996, **251:**245-251.
23. Savchenko A, Charlier D, Dion M, Weigel P, Hallet JN, Holtham C, Baumberg S, Glansdorff N, Sakanyan V: **The arginine operon of *Bacillus stearothermophilus*: characterization of the control region and its interaction with the heterologous *B. subtilis* arginine repressor.** *Mol Gen Genet* 1996, **252:**69-78.
24. Dion M, Charlier D, Wang H, Gigot D, Savchenko A, Hallet JN, Glansdorff N, Sakanyan V: **The highly thermostable arginine repressor of *Bacillus stearothermophilus*: gene cloning and repressor-operator interactions.** *Mol Microbiol* 1997, **25:**385-398.
25. Lu CD, Abdelal AT: **Role of ArgR in activation of the *ast* operon, encoding enzymes of the arginine succinyltransferase pathway in *Salmonella typhimurium*.** *J Bacteriol* 1999, **181:**1934-1938.
26. Xu Y, Liang Z, Legrain C, Ruger HJ, Glansdorff N: **Evolution of arginine biosynthesis in the bacterial domain: novel gene-enzyme relationships from psychrophilic *Moritella* strains (*Vibrionaceae*) and evolutionary significance of N-alpha-acetyl ornithinase.** *J Bacteriol* 2000, **182:**1609-1615.
27. Mironov AA, Koonin EV, Roytberg MA, Gelfand MS: **Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes.** *Nucleic Acids Res* 1999, **27:**2981-2989.
28. Gelfand MS, Koonin EV, Mironov AA: **Prediction of transcription regulatory sites in Archaea by a comparative genomic approach.** *Nucleic Acids Res* 2000, **28:**695-705.

29. Dimova D, Weigel P, Takahashi M, Marc F, Van Duyne GD, Sakanyan V: **Thermostability, oligomerization and DNA-binding properties of the regulatory protein ArgR from the hyperthermophilic bacterium *Thermotoga neapolitana*.** *Mol Gen Genet* 2000, **263:**119-130.
30. Higgins CF, Ames GF: **Two periplasmic transport proteins which interact with a common membrane receptor show extensive homology: complete nucleotide sequences.** *Proc Natl Acad Sci USA* 1981, **78:**6038-6042.
31. Nishijyo T, Park SM, Lu CD, Itoh Y, Abdelal AT: **Molecular characterization and regulation of an operon encoding a system for transport of arginine and ornithine and the ArgR regulatory protein in *Pseudomonas aeruginosa*.** *J Bacteriol* 1998, **180:**5559-5566.
32. Park SM, Lu CD, Abdelal AT: **Cloning and characterization of *argR*, a gene that participates in regulation of arginine biosynthesis and catabolism in *Pseudomonas aeruginosa* PAO1.** *J Bacteriol* 1997, **179:**5300-5308.
33. Berg OG: **Selection of DNA binding sites by regulatory proteins. Functional specificity and pseudosite competition.** *J Biomol Struct Dyn* 1988, **6:**275-297.
34. Summers D, Yaish S, Archer J, and Sherratt D: **Multimer resolution systems of ColE1 and ColK: localisation of the crossover site.** *Mol Gen Genet* 1985, **201:**334-338.
35. Overbeek R, Fonstein M, D'Souza M, Pusch GD, and Maltsev N: **The use of gene clusters to infer functional coupling.** *Proc Natl Acad Sci USA* 1999, **96:**2896-2901.
36. Dandekar T, Snel B, Huynen M, Bork P: **Conservation of gene order: a fingerprint of proteins that physically interact.** *Trends Biochem Sci* 1998, **23:**324-328.
37. **GenBank** [http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Genome]
38. **Genome Therapeutics Corporation** [http://www.genomecorp.com/]
39. Felsenstein J: **Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods.** *Methods Enzymol* 1996, **266:**418-427.
40. **Structural classification of proteins** [http://scop.mrc-lmb.cam.ac.uk/scop/data/scop.1.001.004.004.html]
41. Fogh RH, Ottleben G, Ruterjans H, Schnarr M, Boelens R, Kaptein R: **Solution structure of the LexA repressor DNA binding domain determined by 1H NMR spectroscopy.** EMBO J 1994, **13:**3936-3944.
42. Van Duyne GD, Ghosh G, Maas WK, Sigler PB: **Structure of the oligomerization and L-arginine binding domain of the arginine repressor of *Escherichia coli*.** *J Mol Biol* 1996, **256:**377-391.
43. Sunnerhagen M, Nilges M, Otting G, Carey J: **Solution structure of the DNA-binding domain and model for the complex of multifunctional hexameric arginine repressor with DNA.** *Nat Struct Biol* 1997, **4:**819-826.
44. Ni J, Sakanyan V, Charlier D, Glansdorff N, Van Duyne GD: **Structure of the arginine repressor from *Bacillus stearothermophilus*.** *Nat Struct Biol* 1999, **6:**427-432.

This article is reprinted from Genome **Biology**

Aims and Scope

Genome **Biology** aims to serve the biological research community as an international forum, both in print and on the web, for the dissemination, discussion and critical review of information about all areas of biology informed by genomic research. Subjects covered include any aspect of molecular, cellular, organismal or population biology studied from a genomic perspective, as well as genomics, proteomics, bioinformatics, genomic methods (including structure prediction), computational biology, sequence analysis (including large-scale and cross-genome analyses), comparative biology and evolution.

Publication of primary research on the web

Genome **Biology** offers a very fast publication schedule whilst maintaining rigorous peer-review. Contributors who are in doubt about the suitability of their paper are welcome to send a presubmission enquiry. The editors will provide an initial response to all presubmission enquiries and submitted articles within one working day and will make every effort to give authors a decision following peer review within four weeks of an article's submission. Articles are published electronically as promptly as possible after they are accepted, and within one month of acceptance. The publication date of each article is the date of publication on the web. **The full-length version of all research articles is available to individuals free of charge on the web.**

Copyright, advertising and subscriptions

Genome **Biology** makes no charge to individual authors or readers of primary research articles, and authors retain copyright of their own articles (Genome **Biology** retains the right to print and distribute articles). The costs of providing peer-review and publishing primary papers is offset in part by the subscription fee charged for access to the review, comment and analysis services provide by Genome **Biology** and in part by advertising. If an advertisement appears on this page, its subject should be viewed as completely independent of the content of the article but it is, in part, helping to offset the cost of making primary research available free of charge.

For further information, please contact the editors: editorial@genomebiology.com