J Mol Evol (2001) 53:172–179
DOI: 10.1007/s002390010207

JOURNAL OF MOLECULAR EVOLUTION

Common History at the Origin of the Position–Function Correlation in Transcriptional Regulators in Archaea and Bacteria

Ernesto Pérez-Rueda,* Julio Collado-Vides

Programa de Biología Molecular Computacional, Centro de Investigación sobre Fijación de Nitrógeno, Universidad Nacional Autónoma de México, Cuernavaca, Morelos 62110, México

Received: 1 September 2000 / Accepted: 21 March 2001

Abstract. Regulatory proteins in *Escherichia coli* with a helix–turn–helix (HTH) DNA binding motif show a position–function correlation such that repressors have this motif predominantly at the N terminus, whereas activators have the motif at the C-terminus extreme. Using this initial collection we identified by sequence comparison the exhaustive set of transcriptional regulators in 17 bacterial and 6 archaeal genomes. This enlarged set shows the same position–function correlation. The main question we address is whether this correlation is the result of common origin in evolution or the result of convergence. Evidence is presented supporting a common history at the origin of this correlation. We show the existence of a supergroup of eight repressor protein families sharing a conserved extended sequence comprising the classic HTH. Two of these repressor families (MarR and AsnC) originated before the divergence of Archaea and Bacteria. They are proposed at the origin of HTH-bearing transcriptional regulators currently present in Bacteria. The group of LysR proteins, with the HTH also at the N terminus, offers a control to the argument, since it shows clearly distinctive structural, functional, and evolutionary properties. This group of activator proteins, suggested to have originated within the Bacteria, has an advantageous gene organization to facilitate its horizontal transfer—used to conquer some Archaea—as well as negative autoregulation convenient for homeostasis, all of which agrees with this being the largest family in Bacteria. These results suggest that if shuffling of motifs occurred in Bacteria, it occurred only early in the history of these proteins, as opposed to what is observed in eukaryotic regulators.

Key words: Origin of helix–turn–helix — Location of DNA binding domain — Supergroup — Transcriptional regulators

Introduction

Regulatory mechanisms developed in all organisms appear to be almost-infinite in number, but the basic principles on which they operate are relatively few (von Hippel 1998). The availability of more than 20 prokaryotic genomes offers the opportunity to enrich evolutionary studies of the bacterial transcription machinery. The helix–turn–helix (HTH) motif is one of the most common DNA binding motifs in proteins that control transcription initiation (Sauer et al. 1982). Previously, we have calculated that about 95% of all transcription factors described so far in prokaryotes utilize the HTH motif to bind their target site at the DNA. The remaining 5% is contributed by other binding motifs such as zinc fingers, helix–loop–helix (HLH), β-sheet antiparallel, and RNA binding motif (Pérez-Rueda and Collado-Vides 2000).

In previous analyses (Pérez-Rueda et al. 1998), we have found a correlation between the position of the HTH motif within the proteins and their regulatory role.

* *Present address:* Facultad de Ciencias, Universidad Autónoma del Estado de Morelos, Cuernavaca, Morelos 62210, México
Correspondence to: Julio Collado-Vides, Avenida Universidad s/n CIFN, Cuernavaca, Morelos 62100, México; E-mail: collado@cifn.unam.mx

Proteins with the HTH in their N terminus are either repressor proteins or members of the LysR family of dual regulators (they activate and repress different promoters). Proteins with the HTH motif at the C terminus correspond to a diverse set of regulatory families, most of them activator proteins (Pérez-Rueda et al. 1998). The main motivation of the current study is to investigate whether this correlation is the result of convergence in evolution or the trace of common ancestry.

An exhaustive search for proteins with the HTH motif was performed in the complete genome sequences of 17 bacterial and 6 archaeal species. Evidence is presented supporting the existence of transcriptional regulators sharing a common origin before the divergence of Archaea and Bacteria. We provide evidence for the existence of one supergroup and one ancient family in terms of sequence comparison; the supergroup is associated with repression as its main regulatory function, while the family is associated with dual activity—repression and activation. This analysis suggests a common origin of the proteins forming the supergroup.

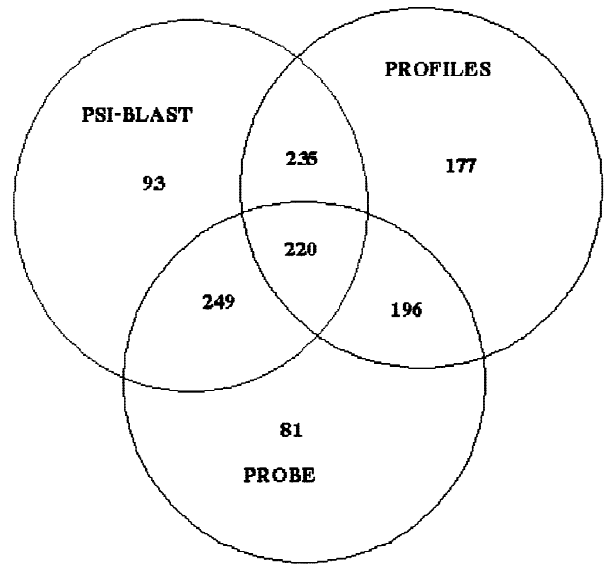

Fig. 1. Venn diagram showing all relationships among methods for predicting transcription factors.

Materials and Methods

Sequence Information

The genome sequences and the ORF annotation for all 23 prokaryotic genomes were acquired from Genbank (ftp://ftp.ncbi.nlm.nih.gov/genbank/Genomes/Bacteria/), except that of *Neisseria meningitidis* strain Z2491, which was acquired from the Sanger Centre (ftp://ftp.sanger.ac.uk/pub/pathogens/), and *Escherichia coli* K-12, which was provided by Blattner et al. (1997).

Identification of Transcription Factors

Three hundred fourteen DNA transcription factors of *E. coli* K12 (Pérez-Rueda and Collado-Vides 2000) were compared against the 42147 predicted proteins of all 23 prokaryotic genomes, using PSI BLAST (cutoff *E* value of 0.05 in the first iteration and 0.01 in the second iteration; BLOSUM 62 probability matrix) (Altschul et al. 1997) to diminish the probability of finding false positives. The outputs were examined using CLUSTAL sequence alignments [with the default conditions (Thompson et al. 1994)] to exclude proteins whose organization is similar to DNA transcription factors but that lack the DNA binding domain (DBD). This first search gave us 872 putative transcription factors in all genomes analyzed. Using the information from the sequence alignments we excluded 75 proteins whose organization lacks the potential HTH, leaving 797 transcription regulators. See Fig. 1.

The PROBE program (Neuwald et al. 1997) was used to detect distant evolutionary relationships among all transcription regulators. PROBE starts a sequence comparison with a sequence "seed" and saves the most similar proteins. The seed sequences used here were the 314 transcription factors described in *E. coli* K12. In a subsequent step it uses the saved sequences to search for additional similar proteins in the database. This process is iterated until it converges with a subset of (distantly) related proteins. Outputs with profiles and homologous protein sequences per iteration are generated which do not necessarily correspond to the DBD. Therefore, we verified the outputs manually to detect additional regulators and to exclude proteins whose organization lack a DBD. From this search, 983 proteins were retrieved. We excluded 237 because they lack the DBD, 469 overlap with the BLAST searches, and 277 are novel putative regulators. In this search, 88 proteins were identified as only one group that corresponds to the LysR family, while 210 proteins were retrieved as a big group in several iterations.

Sixty HTH profiles were built using the Gribskov method (Gribskov et al. 1989, 1987), corresponding to several known regulatory families. These profiles were constructed using all protein families described in *E. coli* K12. One profile corresponds to one family, but additional profiles including more than one family were built when their DBDs in *E. coli* are very similar. This profile combination was originally derived from the high similarity in the DBD in some families in *E. coli,* for instance, the ArsR, AsnC, and GntR families. Operationally we consider as a family proteins sharing at least 25% sequence identity on their DBDs, the criterion also used by Pérez-Rueda and Collado-Vides (2000). A family is a group of proteins that are considered to share a common ancestry (Henikoff et al. 1997). Eight hundred twenty-eight proteins were predicted using this approach. Six hundred fifty-one proteins overlap with BLAST and PROBE, and 177 are exclusive of this procedure.

Alternatively, to corroborate our predictions, hidden Markov model (HMM) profiles were designed with the alignments derived from protein families, using HMMER 2.1.1 (http://hmmer.wustl.edu/). Briefly the method was as follows. Twenty-four HMM profiles were derived from several known protein families in *E. coli,* such as the Crp, GalR/LacI, ArsR, and GntR families (Pérez-Rueda and Collado-Vides 2000). These families were expanded using a larger collection of their orthologue retrieved from Swiss Prot database release 34.0. These proteins were aligned with CLUSTAL W [with the default conditions (Thompson et al. 1994)]. Using this data set of aligned proteins, we built HMM profiles to identify proteins with similar features in the 23 proteomes. In all searches we used as the cutoff an *E* value of 10^{-3} for inclusion as transcription factors. Using this approach, we were able to detect 1075 proteins predicted by the methods described previously (data not shown).

Finally, GenTHREADER, a program that identifies three-dimensional folds in proteins was used (parameters by default), to identify and corroborate the possible fold of the domain corresponding to the HTH motif identified by PROBE (Jones 1999). The set of known regulators of *E. coli* K12 was used as a control set to determine the

accuracy of the methods. Note that we have excluded from the analysis HTH motifs present in RNA polymerase σ factors.

Search results and suplementary material can be accessed at http://www.cifn.unam.mx/Computational Biology/hthpredict.

Results

Data Set of Predicted Regulators

One thousand two hundred fifty-one DNA-binding transcription factors (Tf's) within the 23 genomes were identified using three independent approaches: BLAST searches, sequence profiles searches, and iterative search by PROBE. These 1251 proteins result from taking all those identified by all methods (Fig. 1). All of these factors fall within one of the 24 known regulatory families based on the criterion of family membership mentioned above.

Function assignment of these regulatory proteins was determined using the family assignation. For instance, if the protein belongs to the LysR family, it is asumed to have a dual activity; if it belongs to GalR/LacI, it is considered to be repressor protein. The total number of activators is 260, and that of repressors is 565 (compared with 92 and 113, respectively, in *E. coli*).

Proteins with β-sheet, RNA binding motifs, zinc fingers, and HLH DNA binding motifs were found in lowest proportions. In archaeal organisms, proteins with zinc-finger motifs have been reported, but in a lower proportion than those with HTH (Aravind and Koonin 1999).

Even if the initial set of regulators used in the predictive methods has a large proportion of *E. coli* factors, there are reasons to consider the data set unbiased. Certainly, the HTH motif is not particularly diverse when comparing proteins from different organisms. Family conservation is found across very remotely related prokaryotes, including archaea and remotely related bacteria (see the next section). PROBE detects proteins with distant homology and defines several profiles with proteins from different organisms. When all regulatory proteins were compared, their DBD sequence identity was 25%. Furthermore, even though studies of transcription and of regulatory proteins have traditionally been concentrated in just a few systems (*E. coli*, λ phage, *Salmonella* sp., *Klebsiella* sp., and a few more), there is an important number of experimentally characterized Tf's of bacteria reported in Swiss Prot. In fact, some families contain more members in bacteria other than *E. coli*. Based on these arguments, in addition to the results shown in the next section, we consider it unlikely that new variants of the same HTH motif and additional structural subfamilies will be defined in less-studied bacterial genomes. This does not mean that Tf's with different DNA binding motifs will not be discovered in some of these organisms in the future.

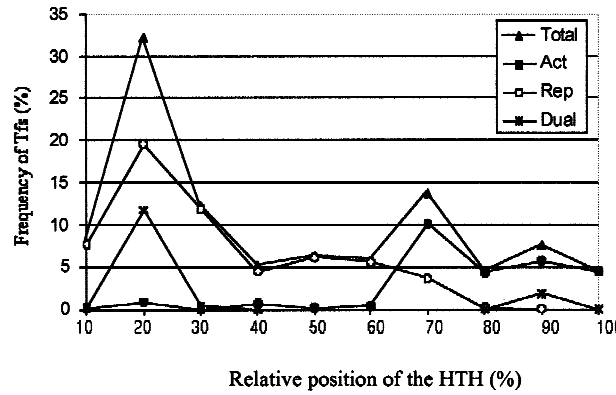

Fig. 2. Distribution of the HTH position in the complete set of 1251 transcriptional regulators. The length of proteins is represented as a percentage, from 0% for the N terminus to 100% for the C terminus. The *Y* axis shows the proportion of proteins at every 10% interval.

One Argument and Two Groups: Common History at the Origin of the Position–Function Correlation in Transcriptional Regulators

Figure 2 shows the distribution of the HTH position in the complete set of 1251 transcriptional regulators, showing the same tendency for repressor proteins to have the DNA binding motif at the N terminus, and C-terminus proteins to be populated mostly by activators. The question we address here is whether this correlation is the result of common origin of groups or the result of convergence in evolution due to physical restrictions related to the activation and repression function.

The Supergroup of Repressor Proteins

In *E. coli* a pattern common to eight regulatory families was detected using the program PROBE—GalR/LacI, DeoR, IclR, GntR, Crp, MarR, MerR, and AsnC (Pérez-Rueda and Collado-Vides 2000)—along with five unclassified proteins (P77484, b0817, YbaE, YmfN, and YabN). This convergence was found using different proteins as the initial "seed," and it is present in about 80% of all known *E. coli* members of these eight families. The pattern is 40 amino acid residues long, located in all cases at the N terminus of the proteins, except in Crp and Fnr, where it is at the C terminus. It contains the core of the HTH motif and a bordering region that in some proteins corresponds to a third α-helix (Suzuki et al. 1995; Wintjens and Rooman 1997). The convergence of the PROBE iterations motivates the proposal of a supergroup of DNA binding Tf's where proteins from different families were clustered, probably sharing a common evolutionary history.

Members of the supergroup share, in addition to sequence conservation, structural and functional properties. They have folds similar to those of the cHTH, such

Table 1. Functional comparison among groups detected[a]

	Regulatory role (%)			
	Repressor	Activator	Dual	Functional feature(s)
Supergroup	55.1	16.4	13.7	C-source uptake, global responses
LysR	1.8	4.4	65.5	Amino acid biosynthesis
All regulators	43	80	20	

[a] Fifty-five and one-tenths percent of all repressor proteins in *E. coli* belong to the supergroup, as well as a small fraction of activators (16.4%) and dual (13.7%) proteins. About 60% of all proteins in the supergroup have a fold similar to that of LacI, Cro, or λ repressors, and 30% similar to that of Crp or DtxR. A few members of the LysR were detected with a fold similar to that of SmtA.

as those of FruR, LacI, PurR, and Cro, except for Crp and their homologues, which have folds similar to those of the wHTH—identified by the fold recognition program GenThreader (Jones 1999) (Table 1). This set groups mostly proteins with repressor function (55% of all repressors described in *E. coli* are included in the supergroup), and they regulate carbon sources uptake. No other proteins within all *E. coli* ORFs were detected by this iterative search procedure. Based on this observation, we hypothesized that if the families clustered in the supergroup are the result of a common origin, the same families would be recuperated when iteratively searching within all prokaryotic genomes.

The iterative search in the 23 prokaryotic genomes shows that, in fact, seven of the eight families found in *E. coli* are recovered sharing the same pattern. In addition to some unclassified proteins, the enlarged number of members of this supergroup belongs to the GalR/LacI, DeoR, IclR, GntR, Crp, MarR, and AsnC families. This pattern (see http://www.cifn.unam.mx/Computational Biology/hthpredict) is similar in position and amino acid composition to the pattern found in *E. coli.* Most of the members of this supergroup are repressors, as suggested by their sequence similarity to characterized repressor proteins. Crp is described as a dual protein, although it has been suggested to have originated as a repressor (Raibaud 1989).

The 1251 regulators analyzed here were grouped in 24 known regulatory families (Pérez-Rueda and Collado-Vides 2000). Figure 3 shows the distribution of repressor protein families and of the LysR family in Bacteria and Archaea. Based on the overall distribution of these families within the 23 microbial genomes, their history can be reconstructed. This reconstruction depends on the ability to distinguish a common origin from horizontal transfer or gene loss. We assume that a consistent frequency of orthologues within a family in an organism, as well as the occurrence of this family within all prokaryotes in a given group, supports a common origin for the family within that group (Lawrence and Ochman 1998). Figure 3 indicates the number of proteins per family in the genomes analyzed. For instance, the ArsR family is clearly present in archaea and several bacteria, whereas MetJ is present only in proteobacteria.

ArsR, MarR, AsnC, and OmpR are suggested ancestor families common to Bacteria and Archaea, given their clear occurrence in most genomes. The first three families are represented in five of the six archaeal genomes in similar or higher proportions compared to their occurrence within the Bacteria. The OmpR family is represented in a smaller amount in only four of them. Although the ArsR family has been associated with mobile elements which could suggest horizontal transfer in Gram-negative and Gram-positive bacteria, such as *E. coli* and *B. subtilis* (Sato and Kobayashi 1998), it is widely present in Archaea and Bacteria. This family contains the shortest repressor proteins, about 100 amino acids long. We have proposed that modules about 100 amino acids long might be the precursors of the DNA binding proteins, because a relatively conserved size of amino acids of proteins within a family has been observed previously (Pérez-Rueda and Collado-Vides 2000). Additional modules with other functions may have given rise to the diversity of transcriptional regulators (Pérez-Rueda et al. 1998). Furthermore, as mentioned above, MarR, as well as the AsnC family, shares the 40-amino acid-long motif common to seven repressor bacterial families defining the supergroup of repressor proteins. Members of the AsnC family involved in DNA organization and amino acid biosynthesis have been postulated to be very ancient, prior to the diversification of the Bacteria, Archaea, and Eucarya (Feng et al. 1997). A member of the AsnC family has recently been shown to bind at multiple sequences in its own promoter in *Sulfolobus solfataricus*, overlapping the TATA box, suggesting negative autoregulation of its own gene (Napoli et al. 1999). This family is, however, absent in *Aquifex eolicus.* Two families of this supergroup, AsnC and MarR, were present before the divergence of Archaea and Bacteria about 3000 million years ago (Feng et al. 1997). We suggest that the HTH of AsnC might be the precursor of the DBD distributed in the supergroup, considering its wide distribution in both domains of life (see Fig. 3). The HTH motif is an ancient nucleotide binding motif that might have been related to the stabilization of the tRNA or rRNA and was later recruited for DNA binding of transcription factors. In fact, it has been argued (Draper 1999) that the recognition strategies and structural frameworks used by RNA binding proteins are not different from those employed by DNA binding proteins, suggesting that the two kinds of nucleic acid binding proteins did not emerge independently in evolution.

The LysR Family: An Independent Group with the HTH at the N Terminus

Members of the LysR family (Henikoff et al. 1988; Schell 1993) were detected as a second group by the

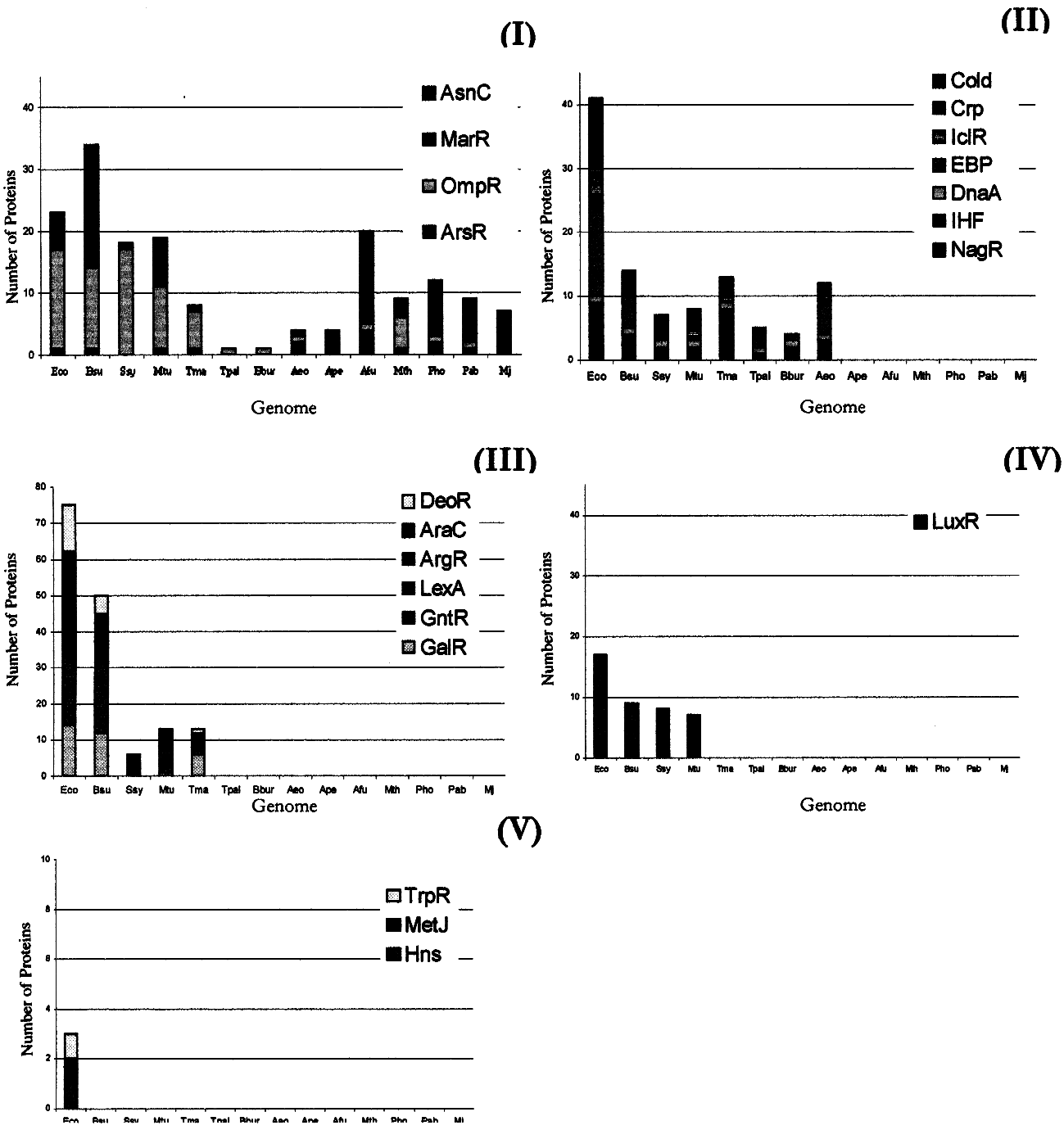

Fig. 3. Distribution of transcriptional factor families in the Bacteria and Archaea. *X* represents the genomes used in the evolutionary inference, and *Y* is the number of proteins per family. **(I)** Families described in all genomes, **(II)** families exclusive to bacteria, **(III)** families described in thermotogales and bacteria that evolved afterward, **(IV)** families associated with firmicutes and proteobacteria, and **(V)** families found only in proteobacteria. Bacterial species: Eco, *Escherichia coli;* Bsu, *Bacillus subtilis;* Ssy, *Synechocystis* sp. *synecho;* Mtu, *Mycobacterium tuberculosis;* Tma, *Thermotoga maritima;* Tpa, *Treponema pallidum;* Bbu, *Borrelia burgdorferi;* Aeo, *Aquifex aeolicus.* Archaeal species: Ape, *Aeropyrum pernix;* Afu, *Archaeaoglobus fulgidus;* Mth, *Methanobacterium thermoautotrophicum;* Pho, *Pyrococcus horikoshii;* Pab, *Pyrococcus abyssi;* Mj; *Methannococcus jannaschi.* Bacteria excluded from the inference: Hpy, *Helicobacter pylori;* jhp, *H. pylori* strain J99; Cpn, *Chlamydia pneumoniae;* Ctr, *Chlamydia trachomatis;* Rpr, *Rickettsia prowazekii;* Mpn, *Mycoplasma pneumoniae;* Mge, *Mycoplasma genitalium;* Rng, *Rhizobium* sp. NGR234; Hin, *Haemophilus influenzae.*

iterative search. This family has the DBD at the N terminus and two additional motifs that might be involved in metabolite recognition or in the multimerization process (Schell 1993).

The importance of this second group is that it clearly separates the two classes of proteins with their HTH in the N terminus into the supergroup of negative regulators and the LysR family of dual regulators, or, more precisely, activators that repress their own expression. Members of this family have been described in almost all bacteria, although in some cases associated with mobile elements that could help to diversify their regulatory function (Schell 1993). This family is represented by two members in *A. fulgidus* and in *M. thermoautotrophicum* and by one member in *M. janashii*, suggesting their horizontal transfer from Bacteria.

The genes for these regulators are usually adjacent to a regulated gene in the opposite direction, defining a divergent pair of regulated promoters. This transcriptional organization would facilitate its horizontal transfer following the selfish operon hypothesis (Ochman et al. 2000), since, in addition to functionally related, closely located genes, it offers a way to transfer the regulatory region (its promoter and operator) within the noncoding sequence, something particularly useful given its autoregulatory function. Furthermore, the negative autoregulation of LysR proteins and its homeostatic effect (Thieffry and Thomas 1995) is one additional argument that might have contributed to its evolutionary great success. Certainly, it is the largest family within the Bacteria.

In brief, members of the LysR family define an evolutionary group clearly distinct from the supergroup of repressors discussed before. They have a different conserved sequence pattern; their role is mostly as activators with negative autoregulation; they must have emerged in Bacteria, where they are the dominant group, with an exceptional presence in the Archaea; and they share an operon organization that seems ideal for horizontal transfer. The existence of this group gives further support to the argument of a common origin for the correlation between the position of the HTH and the role of activators in transcription, as opposed to convergence due to physical restrictions.

Discussion

It is not possible to know in detail the historical events at the origin of transcription factors, because the small size of the motif is not informative enough to reveal whether all HTH motifs come from a unique ancestor in evolution (Rosinski and Atchley 1999). However, as illustrated here, it is possible to know whether a group of families sharing some properties is the result of a common origin or the result of convergence in evolution.

No other bacterial family in addition to the repressor families discussed here is represented within a single archaeal genome. Previous studies have suggested that the HTH motif of specific Tf's occurs in Archaea as a result of horizontal transfer from Bacteria (Aravind and Koonin 1999; Koonin et al. 1997; Makarova, et al. 1999). Our results indicate that the HTH motif of specific Tf's occurs in Bacteria and Archaea as a result of a common origin, except for the few cases just discussed. The co-occurrence of some families within proteins sharing a common pattern beyond the HTH provides additional evidence for a common origin of this type motif before the divergence of Archaea and Bacteria.

It is interesting in this sense that these ancient families are basically repressors of the σ70 promoters in bacteria, and they may as well act as repressors within the archaeal transcription machinery. The archaeal basal transcription apparatus shows a striking similarity to the eukaryotic machinery such as the TATA box, RNA polymerase, and transcription factors (Bell and Jackson 2000). One may wonder if this is in apparent contradiction with the dominance of activation in eukaroytes. However, the ability of in vitro transcription with reconstituted TBP, TFB, and RNA polymerase suggests the possible presence of basal transcription activity of promoters in vivo and, therefore, the need of the cell to repress them. Recent evidence shows that in *A. fulgidus* a promoter is subject to repression by preventing recruitment of the RNA polymerase by a protein similar to the DtxR bacterial repressor (Bell et al. 1999). This adds to the plausible repression of an Lrp-like regulator in *Sulfolobus* (Napoli et al. 1999).

Horizontal transfer among bacteria, archaea, and eukaryotes has been suggested to occur frequently (Brown and Doolittle 1997), and although genes whose products are involved in protein–protein interactions are expected to be transferred less efficiently (Jain et al. 1999), the conservation of the RNA polymerase in bacteria (Gruber and Bryant 1997), and the binding of several activators to the same region of the RNA polymerase (Lonetto et al. 1998), should eliminate this restriction for the activator or repressor proteins to be transferred within the Bacteria.

Two major groups of transcription regulators can be identified based on the HTH positioning in the protein, the N-terminus and the C-terminus groups, which correlate with repressors and activators, respectively. Evidence presented in this paper indicates that the majority of repressor proteins shares a common origin. This evidence includes the positional conservation of the HTH, the presence of important amino acids involved in the HTH conformation, sequence conservation beyond the strict HTH motif, and the clear occurrence of some of these repressors before the Archaea and Bacteria divergence. A possible scenario for the evolution of transcriptional HTH regulators could start with a relatively small repressor as ancestor to the actual Tf's, before the divergence of Archaea and Bacteria. Afterward, the joining of

larger domains involved in metabolite recognition, or multimerization, determined the HTH DNA binding domain within the sequence: as a consequence of this joining, it would be near the end N or C terminus of the protein. An ancestral protein with a DNA domain located N-terminal to the ligand domain could have originated the repressors. Very ancient duplications eventually gave rise to the different members of the supergroup. Similar processes of domain addition have been proposed for HTH modules (Jones, 1999), in Two-Component proteins (Pao and Saier 1995) and in other families, such as GalR/LacI and NagR/XylR (Titgemeyer et al. 1994; Weickert and Adhya 1992).

Although the protein sequence has changed significantly within the families clustered in the supergroup, the variation in the protein backbone has been smaller within the HTH structure (Wintjens and Rooman 1997). This smaller variation would reflect the functional conservation among several members of the supergroup.

Topological restrictions on multimerization could be related to the positioning of the motif at the N terminus (Gralla and Collado-Vides 1996) for repressors. In fact Percipalle et al. (1995) have proposed that positioning of the HTH in the N terminus contributes to greater flexibility in a dimer, promoting stronger binding to DNA, implying a better role as a repressor. Its counterpart, positioning of the HTH in the N terminus, does not prevent protein to activation as illustrated by the LysR family. These dual regulators repress their own gene by binding to a located site closer to the binding of the RNA polymerase than its usual activator position (Gralla and Collado-Vides 1996). Overall this evidence suggests that the better explanation for the existence of the supergroup of repressors is a common origin, and not functional restriction.

Acknowledgments. E.P.R. was supported by a doctoral fellowship from CONACYT and DGEP-UNAM. Part of this work was supported by grants from CONACYT (No. 0028) and from DGAPA to J.C.V. We acknowledge important discussions with Alejandro Garciarubio, Arturo Medrano-Soto, and Enrique Morett, as well as suggestions on a previous version of the manuscript by Gabriel Moreno-Hagelsieb, Antonio Lazcano-Araujo, and anonymous referees.

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller, W, Lipman D (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs Nucleic Acids Res 25:3389–3402

Aravind L, Koonin EV (1999) DNA-binding proteins and evolution of transcription regulation in the archaea. Nucleic Acids Res 27:4658–4670

Bell SD, Jackson SP (2000) Mechanism of autoregulation by an archaeal transcriptional repressor. J Biol Chem 275:12934–12940

Bell SD, Cairns SS, Robson RL, Jackson SP (1999) Transcriptional regulation of an archaeal operon in vivo and in vitro. Mol Cell 4:971–982

Blattner FR, Plunkett G III, Bloch CA, et al. (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277:1453–1462

Brown JR, Doolittle WF (1997) Archaea and the prokaryote-to-eukaryote transition Microbiol. Mol Biol Rev 61:456–502

Draper DE (1999) Themes in RNA-protein recognition. J Mol Biol 293:255–270

Feng DF, Cho G, Doolittle RF (1997) Determining divergence times with a protein clock: Update and reevaluation. Proc Natl Acad Sci USA 94:13028–13033

Gralla JD, Collado-Vides J (1996) Organization and function of transcriptional regulatory elements. In: Neidhart FC, Curtiss R III, Ingraham JL, (eds) *Escherichia coli* and *Salmonella typhimurium:* Cellular and molecular biology. Am Soc Microbiol, Washington, DC, pp 1232–1246

Gribskov M, McLachlan AD, Eisenberg D (1987) Profile analysis: Detection of distantly related proteins. Proc Natl Acad Sci USA 84:4355–4358

Gribskov M, Luethy R, Eisenberg D (1989) Profile analysis. In: Methods in enzymology. Academic Press, San Diego, pp 146–159

Gruber TM, Bryant DA (1997) Molecular systematic studies of eubacteria, using sigma70-type sigma factors of group 1 and group 2. J Bacteriol 179:1734–1747

Henikoff S, Haughn GW, Calvo JM, Wallace JC (1988) A large family of bacterial activator proteins. Proc Natl Acad Sci USA 85:6602–6606

Henikoff S, Greene EA, Pietrokovski S, Bork P, Attwood, TK, Hood L (1997) Gene families: The taxonomy of protein paralogs and chimeras. Science 278:609–614

Jain R, Rivera MC. Lake JA (1999) Horizontal gene transfer among genomes: The complexity hypothesis. Proc Natl Acad Sci USA 96:3801–3806

Jones DT (1999) GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences J Mol Biol 287: 797–815

Koonin EV, Mushegian AR, Galperin MY, Walker DR (1997) Comparison of archaeal and bacterial genomes: Computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. Mol Microbiol 25:619–637

Lawrence JG, Ochman H (1998) Molecular archaeology of the *Escherichia coli* genome. Proc Natl Acad Sci USA 95:9413–9417

Lonetto MA, Rhodius V, Lamberg K, Kiley P, Busby S, Gross C (1998) dentification of a contact site for different transcription activators in region 4 of the *Escherichia coli* RNA polymerase sigma70 subunit. J Mol Biol 284:1353–1365

Makarova KS, Aravind L, Galperin MY, Grishin NV, Tatusov RL, Wolf, YI. Koonin EV (1999) Comparative genomics of the Archaea (Euryarchaeota): Evolution of conserved protein families, the stable core, and the variable shell. Genome Res 9:608–628

Napoli A, Van der Oost J, Sensen CW, Charlebois RL, Rossi M, Ciaramella M (1999) An Lrp-like protein of the hyperthermophilic archaeon *Sulfolobus solfataricus* which binds to its own promoter. J Bacteriol 181:1474–1480

Neuwald AF, Liu JS, Lipman DJ, Lawrence CE (1997) Extracting protein alignment models from the sequence database. Nucleic Acids Res 25:1665–1677

Ochman H, Lawrence JG, Grolsman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304

Pao GM, Saier MH Jr (1995) Selective domain shuffling during evolution J Mol Evol 40:136–154

Percipalle P, Simoncsits A, Zakhariev S, Guarnaccia C, Sanchez R, Pongor S (1995) Rationally designed helix-turn-helix proteins and their conformational changes upon DNA binding. EMBO J 14: 3200–3205

Pérez-Rueda E, Collado-Vides J (2000) The repertoire of DNA-binding transcriptional regulators in *Escherichia coli* K12. Nucleic Acids Res 28:1838–1847

Pérez-Rueda E, Gralla J, Collado-Vides J (1998) Genomic position analyses and the transcription machinery. J Mol Biol 275:165–170

Raibaud O (1989) Nucleoprotein structures at positively regulated bacterial promoters: Homology with replication origins and some hypotheses on the quaternary structure of the activator proteins in these complexes. Mol Microbiol 3:455–458

Rosinski JA, Atchley WR (1999) Molecular evolution of helix-turn-helix proteins. J Mol Evol 49:301–309

Sato T, Kobayashi Y (1998) The ars operon in the skin element of *Bacillus subtilis* confers resistance to arsenate and arsenite. J Bacteriol 180:1655–1661

Sauer RT, Yocum RR, Doolittle RF, Lewis M, Pabo CO (1982) Homology among DNA-binding proteins suggests use of a conserved super-secondary structure. Nature 298:447–451

Schell MA (1993) Molecular biology of the LysR family of transcriptional regulators. Annu Rev Microbiol 47:597–626

Suzuki M, Yagi N, Gerstein M (1995) DNA recognition and superstructure formation by helix-turn-helix proteins. Prot Eng 8:329–338

Thieffry D, Thomas R (1995) Dynamical behaviour of biological regulatory networks. II. Immunity control in bacteriophage lambda. Bull Math Biol 57:277–297

Thompson JD, Higgins DG, Gibson TJ (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Titgemeyer F, Reizer J, Reizer A, Saier MH Jr (1994) Evolutionary relationships between sugar kinases and transcriptional repressors in bacteria. Microbiology 140:2349–2354

von Hippel PH (1998) An integrated model of the transcription complex in elongation, termination, and editing. Science 281:660–665

Weickert MJ, Adhya S (1992) A family of bacterial regulators homologous to Gal and Lac repressors. J Biol Chem 267:15869–15874

Wintjens R, Rooman M (1997) Structural classification of HTH DNA-binding domains and protein-DNA interaction modes. J Mol Biol 262:294–313