Conservation of transcriptional sensing systems in prokaryotes: A perspective from Escherichia coli

Heladia Salgado\textsuperscript{a}, Agustino Martínez-Antoniob,\textsuperscript{a,*}, Sarath Chandra Janga\textsuperscript{a,}\textsuperscript{*}

\textsuperscript{a} Programa de Genómica Computacional, Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, Morelos 62100, Mexico

\textsuperscript{b} Departamento de Ingeniería Genética, Centro de Investigación y de Estudios Avanzados, Instituto Politécnico Nacional, Campus Guanajuato, Irapuato 36500, Mexico

Received 7 May 2007; revised 14 June 2007; accepted 22 June 2007
Available online 2 July 2007
Edited by Robert B. Russell

Abstract The activity of transcription factors is usually governed by allosteric physicochemical signals or metabolites, which are in turn produced in the cell or obtained from the environment by the activity of the products of effector genes. Previously, we identified a collection of more than 110 transcription factors and their corresponding effector genes in Escherichia coli K-12. Here, we introduce the notion of “triferog”, which relates to the identification of orthologous transcription factors and effector genes across genomes and show that transcriptional sensing systems known in E. coli are poorly conserved beyond Salmonella. We also find that enzymes that act as effector genes for the production of endogenous effector metabolites are more conserved than their corresponding effector genes encoding for transport and two-component systems for sensing exogenous signals. Finally, we observe that on an evolutionary scale enzymes are more conserved than their respective TFs, suggesting a homogenous cellular metabolism across genomes and the conservation of transcriptional control of critical cellular processes like DNA replication by a common endogenous signal. We hypothesize that extensive variation in the domain architecture of TFs and changes in endogenous conditions at large phylogenetic distances could be the major contributing factors for the observed differential conservation of TFs and their corresponding effector genes encoding for enzymes, causing variations in transcriptional responses across organisms.

© 2007 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.

Keywords: Transcriptional sensing system; Cellular signaling; Conservation; Transcriptional regulation; Allosteric regulation; Effector metabolite; Prokaryotes

1. Introduction

Organisms constantly monitor environmental conditions in order to respond to changes. This is normally achieved by physicochemical signals, which are recognized by the cell as messengers of environmental composition or their own metabolic state [1–3]. The binding of transcription factors (TFs) to these specific signals determines their active or inactive conformation and affinities to interact with specific sequences on cis-regulatory regions of transcription units or with the rest of the transcriptional machinery [4]. In turn these signals are produced or delivered as a result of the activity of the protein products of the effector genes, which indirectly but effectively control the activity of the TFs, providing a concerted link between the genetic and metabolic components of a cell in the regulation of transcription. Recent studies demonstrated that the components of the network of transcriptional interactions (i.e. a network where TFs and their regulated genes form the nodes and the directed regulatory interactions form the edges) in bacterial genomes evolve rapidly [5,6]. However, it is unclear if the genetic repertoire forming the core of the transcriptional regulation in response to allosteric signals follows the same trend of poor conservation across organisms. In this work, we study the conservation of the genetic components for transcriptional sensing systems (TSSs), that is to say, the TFs and their corresponding effector genes, across a range of prokaryotic organisms and show that the TSSs identified in the gram negative γ-proteobacterium, E. coli K-12 [7–9], are poorly conserved across the phylogenetic spectrum, with closely related species sharing a higher proportion of the TF-effector pairs. We find that despite poor conservation of TSSs across genomes, certain TF-effector gene pairs sensing basic metabolites like ATP, biotin, amino acids and some metals are highly conserved. Our observations suggest that TSSs behave like functional modules as their component TF and signal genes were found to be significantly co-detected across a set of non-redundant genomes. Furthermore, we also demonstrate that there is a variation in the extent of conservation of different transcriptional signal sensing categories as defined by the effector genes constituting them. The results reported here should enhance our understanding of the evolution of transcriptional sensing machinery in prokaryotes.

2. Materials and methods

2.1. Dataset

Information about TFs and their corresponding effector genes with experimental evidence from literature was gathered from the RegulonDB database, which contains extensive information centered on transcriptional regulation in E. coli strain K-12 [10]. Our initial dataset comprised 84 TFs and 291 corresponding effector genes as described in...
Martinez-Antonio et al. [7]. It should be noted that a TF to effector gene relation can be many-to-many, so a TF can have multiple effector genes and multiple TFs can be controlled by a single effector gene. In RegulonDB, however, information on their signal effectors is available for only a minor fraction of all known two-component systems which are experimentally characterized. Therefore, we added to this list those cases for which there was indirect evidence (such as their experimental characterization in closely related species). Hence, our final dataset consisted of 113 TFs covering around 38% of the roughly 297 predicted TFs in E. coli [7,11,12]. The complete classification of TFs and effector genes is available as Supplementary material.

2.2. Identification of orthologous transcription factors and effector genes across genomes
Orthologs are defined as those genes in different species that evolved from a common ancestor by speciation [13] and usually have the same function. Our working definition of orthology consisted of BLASTP reciprocal best hits, along with additional rigorous parameters to take into account the multi-domain nature and extensive duplication in transcription factors [11] as described earlier [6]. The majority of TFs in prokaryota are comprised of at least a DNA-binding domain and an effector domain. In order to identify and distinguish orthologous sequences from those arising due to lineage specific duplications and recombinations, which are known to be the common phenomena driving the evolution of TFs [11,14], it becomes important to consider domain organization and orientation to detect functionally equivalent orthologs. Therefore in addition to traditional bi-directional best hits it was required that Pfam domains [15] of query and target proteins match to consider them as orthologous. Using this approach we identified the orthologs of all protein coding genes in E. coli across a collection of 216 genomes shown in the Supplementary material. More relaxed definitions of orthology which rely only on reciprocal best hits and typically gain on the coverage of orthology detection but loose on sensitivity [5] did not vary our conclusions (see Supplementary material).

2.3. Validation of detected triferogs
In order to assess how likely triferogs are to be detected in a genome due to chance alone and to measure the significance of their conservation, we compared the number of triferogs identified in a given genome for the complete set of TSSs in E. coli against the conservation of 1000 randomly constructed TSS collections. Each random collection was created by randomly assigning an effector gene to a TF by simply altering the label of the effector gene with any protein coding gene in E. coli except that of the TF itself, while the label of the TF was retained as such. In order to avoid over-scoring on the extent of conservation due to the over-representation of genomes which are evolutionarily very close, we filtered out strains and species of the same bacterial genus, keeping the strain or species with the maximum number of genes among a given genus of organisms to generate a non-redundant set of genomes as described earlier [6]. In addition, to avoid any affects due to the sample size of conserved pairs, we considered only those non-redundant genomes in which at least 20 triferogs were identified for presenting our results. P-values were calculated from Z-scores assuming a normal distribution of the random observations since the number of conserved pairs for the random TSS datasets followed a gaussian distribution.

3. Results and discussion

3.1. Triferog definition
A transcriptional sensing system (TSS) is comprised of three elements: (i) a transcription factor protein, (ii) an effector gene and, (iii) a corresponding signal effector. These systems may or may not constitute transcriptional sensing circuits, i.e. the effector gene may or may not be directly regulated by its corresponding TF (see Fig. 1a and b). The effector gene may not be directly regulated by its corresponding TF when it is involved in a regulatory cascade of two or more TFs where an effector gene regulated by one TF could be producing an effector metabolite modulating the activity of a second TF (see Fig. 1b). Apart from this, there could be a possibility that the first and second metabolites correspond to higher and lower parts of the same metabolic pathway. In this way the transcriptional sensing systems play important roles to link the transcriptional regulation of genes whose products are involved in different parts of metabolism or a regulatory cascade and shape the cell physiology to varying exogenous and endogenous conditions [2]. In the whole network of TF-effector gene pairs, we found that only ~36% of the links could be accounted for transcriptional interactions between a TF and its effector gene, suggesting that majority of the effector genes are not under the transcriptional control of their respective TFs but control the activity of the TFs in an indirect manner (see Supplementary material). In this work, we introduce the notion of ”triferog”, which refers to the presence of both an orthologous transcription factor and its effector gene in different bacteria. If these two genetic components are present in distant bacteria it is probable that they constitute a TSS involving the same signal effector as seen in the reference or source genome (see Fig. 1c). The process of identifying a putative triferog in an organism of interest involves the assignment of an orthologous effector gene to its corresponding TF which is likely to control or modulate its activity. This concept is important to understand the conservation of transcriptional sensing machinery across bacteria. It should be noted that the concept of triferog is different from that of regulon, while in the former a TF and its effector gene constituting a regulatory interaction, is more of an exception than a rule as this happens in less than 33% of the regulons (see Supplementary material).

3.2. Transcriptional sensing systems identified in E. coli are poorly conserved across γ-proteobacterial genomes
To study the conservation of TSSs in E. coli we identified their triferogs across a collection of 216 completely sequenced genomes tabulated in the Supplementary material. In Fig. 2a, we show the proportion of TSSs conserved across all the γ-proteobacterial genomes from this set. As the phylogenetic distance with respect to E. coli K-12 increases, the proportion of triferogs identified across genomes decreases. All the 5 strains of E. coli share more than 80% of the TSSs known in E. coli K-12, while all the Shigella and Salmonella species share between 70% and 80%. All Yersinia strains and the Pectobacterium, Erwinia carotovora share between 50% and 60% of the TSSs known in E. coli. Vibrionaceae, Photobacterium luminescens and Pseudomonas aeruginosa were found to share between 35% and 50% while Pasteurellaceae, which include Haemophilus influenzae, Pasteurella multocida and Shewanella species have around 20–35% of the TSSs conserved, indicating that the genetic components composing the TSSs in E. coli K-12 are poorly conserved beyond Salmonella and Shigella genomes.

3.3. Conservation of the genetic machinery for sensing drops rapidly as the phylogenetic distance with respect to E. coli increases
Fig. 2b shows the conservation of TSSs across 101 non-redundant set of genomes (see Section 2 and Ref. [6], e.g. only
**E. coli** in this example, CysB is the master regulator for sulfur metabolism in its respective TF but can still respond to a functionally related TF. In (b) show how the effector gene may or may not be directly regulated by ancient duplicate of cysB TSS is experimentally shown to be conserved [32,33].

Transcriptional sensing system (TSS) observed in **E. coli**

Fig. 1. (a) An example showing the mechanism of action of a transcriptional sensing system (TSS) observed in **E. coli**. The effector gene cyaA encodes for the enzyme adenylate cyclase that catalyzes the formation from cyclic-AMP using ATP as substrate in the cytoplasm. The cAMP acts as signal metabolite (mainly in carbon source starvation) and when it is bound by the cAMP receptor protein (CRP), its transcriptional activity is affected by allosterism. This CRP–cAMP complex regulates the transcription of many transcription units (CRP), its transcriptional activity is affected by allosterism. This CRP–cAMP complex regulates the transcription of many transcription units.

(b) Representation of two interconnected TSSs involved in the regulation of sulphur assimilation in **E. coli** K-12. CysB and Cbl are controlled by their corresponding signal metabolites produced by the effector genes for transport and synthesis. CysB acts depending on the availability of transportable metabolites containing sulfur (in this case thiolsulphate is the signal metabolite) and Cbl responds to the signal metabolite (adenosine 5-phosphate) which is a product of the initial biochemical steps in the synthesis of L-cysteine. These TSSs in (a) and (b) show how the effector gene may or may not be directly regulated by its respective TF but can still respond to a functionally related TF. In this example, CysB is the master regulator for sulfur metabolism in **E. coli** and it is possible that cbl, an accessory regulatory partner, is an ancient duplicate of cysB as these genes share 60% amino acid identity [30,31].

(c) Representation of the Triferog concept using the TSS shown in (a) for **E. coli** K12. As the genetic sensing components in the source organism are identified and validated as orthologs (dashed lines) in a different genome (target organism) it is highly probable that this genetic sensing system might be responding to the same signal effector. The black lines represent the signal effector availability by the effector(s) gene(s) and their allosteric effect on the respective transcription factor in the source organism. The break lines in the target organism represent the putative TF-effector interaction between the orthologs of genetic sensing components observed in the source organism. In this case the target organism is **Vibrio cholerae** where the TSS is experimentally shown to be conserved [32,33].

3.4. **TF-effector pairs constituting enzymes as effector genes are more conserved than those comprising transporters and sensor proteins**

An obvious question which arises, given that TSSs identified in **E. coli** are poorly conserved across genomes with increasing phylogenetic distance is: Are there TSSs or their components which are evolutionarily more conserved and hence are likely to be ancient and more stable, and if so, what are their functions? To address this, we analyzed the conservation of TSSs by taking into account the type of effector genes constituting them. We were able to identify only three major classes of TSSs, depending on the effector genes constituting them, namely enzymes, transporters and sensor proteins. However, there might be some poorly represented or unidentified class of effector genes in the complete regulatory repertoire of the cell which we can not take into consideration at the moment due to the incompleteness of our knowledge about transcriptional regulation in **E. coli** K12. The first kind of the three classes that we consider in this study includes TFs sensing signal metabolites synthesized by enzymes in the cell cytoplasm and therefore might correspond to the sensing of intracellular signals while the other two comprise of TFs which transporters and sensor proteins (of two-component systems) act as their effector genes and sense mostly exogenous signals [7,8].

Fig. 3 shows the conservation of TSSs based on this classification. There is a clear bias in the extent of conservation of TSSs constituting enzyme genes (i.e. preponderance of triferogs of internal sensing systems identified in **E. coli** in comparison to those comprising transporter and sensor proteins (see
Fig. 3a, b and c, respectively, the X-axis shows the conservation of TF and effector gene pairs across bacterial genomes. This finding makes biological sense because signals sensed by well conserved TSSs of the enzyme class correspond to important building block metabolites involved in cell structure, intermediate metabolic pathways or as cellular fuel. The most conserved TSS with its trifereg detected in 130 genomes is that constituted by DnaA (DNA-replication initiator protein) and...
some of the enzymes involved in the ATP (adenosine triphosphate) biosynthesis, followed by those corresponding to the synthesis of arginine (ArgR), biotin (BirA), glycerol-3-phosphate (GlpR) and leucine (Lrp), ranking from 71 to 58 genomes in their extent of conservation (see Figs. 4 and 3a). It is important to note that arginine, leucine and glycerol-3-phosphate are effector signals of hybrid TSSs in *E. coli* as they can be produced both endogenously (using enzymes as effector components).
genes) and exogenously (using transporters as effector genes) [7]. The cell, in addition to having enzymes necessary for synthesizing these metabolites internally, can also import them from the exterior of the cell using transport systems. This means that certain TFs have effector genes encoding for both enzymes and transporters and therefore are capable of sensing both exogenous and endogenous conditions. The transport systems used for obtaining these metabolites from the milieu are less conserved across bacteria than their corresponding enzymes (see Fig. 4) and this bias is observed in other hybrid systems (see Supplementary data of sensing systems and Fig. 4).

Among the TSSs involving the transport of exogenous metabolites the most conserved are those for internalizing metals like zinc and ferric ion (in 70 and 51 genomes, respectively) while in two-component systems highly conserved pairs occur in less than 50 genomes (Fig. 3b and c). Taken together, these observations allow us to conclude that TSSs for internal signals in *E. coli* are more conserved in bacteria in comparison to those for sensing exogenous signals. This becomes very evident from those cases where the cell, despite having the machinery for transporting the metabolites from the milieu in addition to synthesizing them internally, prefers to conserve biosynthesis systems over transport systems. This might be due to the fact that certain internal signals are also important metabolites for the cellular metabolism and hence can be expected to be more homogeneously distributed across different bacterial kingdoms in comparison to the external signals, which change depending on the composition of each biological niche. As a
consequence the transcriptional machinery to detect the external signals might be niche-dependent.

3.5. Transcription factors and their effector genes are conserved across genomes to unequal extents when compared against the conservation of their TSSs

It is well known that different cellular components are differentially conserved across the phylogenetic tree. In the case of transcriptional regulatory networks it has been shown recently that the transcription factors evolve faster than their target genes [5,6]. In a similar way it is possible to analyze separately the genomic conservation of transcription factor and effector genes. Thus in the Y-axis of Fig. 3 conservation of the components of the TSSs in bacterial genomes is shown. It is easy to note that enzymes are more conserved than their TFs while transporters and their respective TFs seem to be conserved to the same extent. On the other hand, sensor components are less conserved than their corresponding response regulators (Fig. 3a, b and c, respectively). It is possible that enzymes are more conserved than other kind of effector genes with respect to their TFs as they form part of the basal cellular metabolism producing important intracellular metabolites as discussed in the previous section. For example, Eno (enolase, conserved in 177 genomes) is the enzyme involved in the interconversion of 2-phosphoglycerate to phosphoenolpyruvate. Following are some well conserved enzymes: GlyA (serine aldolase subunit, 172 genomes) involved in the interconversion of glycine and serine, some components of ATP synthesis (like AtpA, 162 genomes), GpsA catalyzing the conversion of dihydroxyacetone-phosphate to produce glycerol-3-phosphate (157 genomes) and MetK (S-adenosylmethionine synthetase, 148 genomes). All these enzymes are more conserved than the most conserved TFs, DnaA and BirA (146 and 86 genomes, respectively). It is interesting to note in the case of TSSs where the effector genes encode for transporters, TFs and their respective effector genes are conserved to equal extents (Fig. 3b), except in the case of Fur (ferric uptake regulator), where the TF is far more conserved than its effector gene for transport. A possible explanation for this observation could be that most of these TSSs are either encoded in the same operon or in close chromosomal proximity which can aid them to evolve as chromosomal modules [9]. On the other hand lower conservation of sensor genes with respect to their response regulators in two-component systems is difficult to explain given that these systems also tend to be encoded proximal on the E. coli chromosome and are found to be interacting even at the level of protein products [9]. One possible reason for this differential conservation of TFs and their cognate sensor genes could be due to the extreme flexibility in the genetic components of two-component systems as a result of horizontal gene transfer events and lineage-specific expansions in bacteria [17]. In addition it is well known that the signal input domain in the histidine kinases is highly variable [18] and therefore detection of orthologs might be effected, causing imbalance in the detection of two-component systems. Although in general there is a tendency for an equilibrium in the extent of conservation of response regulators and histidine kinases in bacteria, there are some groups where there are clearly more histidine kinases than response regulators as seen in cyanobacteria and green sulfur bacteria and on the contrary beta and epsilon-proteobacteria have less histidine kinases than response regulators [19,20]. Thus it might be that sensor components of the external sensing machinery vary more quickly in bacteria given that some redundancy and cross talk in some of their components is known to exist [21,22]. Thus it seems that much of the variation in transcriptional sensing machinery across bacteria can be accounted for the changes in the effector genes used for sensing external conditions, which are poorly conserved, than those for sensing the changes in the endogenous conditions.

In this work, we show that TSSs identified in E. coli, which can be interpreted to comprise of TF-effector gene pairs, are poorly conserved across genomes and the conservation falls off rapidly with increasing phylogenetic distance. We also find that TSSs from different categories are conserved to varying extents in complete genomes, with those comprising enzymes as their effector genes conserved the most. Our results suggest that transcriptional sensing machinery involved in the sensing of signals mostly synthesized by enzymes in the cytoplasm is well conserved across organisms, while the one sensing exogenous conditions is poorly shared with phylogenetically distant organisms. We note that some of the TSSs shown in Fig. 4 (mainly those constituted by BirA, Lrp, GlpR, ArgR and Fur) in E. coli form transcriptional sensing circuits, where the TF regulates its effector gene (note that less than 35% of the TFs show this property). However, due to the lack of information about transcriptional regulation in phylogenetically distant organisms it would be premature to conclude if these kind of circuits would tend to be more conserved than those TSSs not forming circuits, across bacterial genomes. The differential conservation of TF and effector gene (enzyme) pairs that we observe in this work can be explained by the following factors: (a) although the cellular metabolism might be conserved, each group of bacteria depending on their life-history could be using different set of endogenous metabolites to control their gene expression and consequently use different sets of effector genes to affect the activity of non-orthologous or different TFs. In fact, even if they use the same endogenous metabolite and hence same effector gene, transcriptional responses could be very different across organisms due to a plethora of possibilities in the domain combinations of TFs. (b) post-transcriptional mechanisms like regulation by riboswitches might be substituting regulation at the level of transcription played by TFs, by responding to identical regulatory signals, as seen in the biosynthesis of the amino acids, methionine and tryptophan, where in the regulation in B. subtilis and E. coli is operated by different means [23–25]. It is possible to hypothesize based on our observations that at large phylogenetic distances there could be extensive variations in the domain architecture of the repertoire of TFs to accommodate and suffice the variations in the endogenous conditions of the cell, despite keeping some of the core enzymatic roles a constant. It would be interesting to explore in greater detail, on a case by case basis across genomes, where the metabolic pathway is known to be conserved but no corresponding TF is detected as this will enhance our knowledge about novel mechanisms linking metabolic and transcriptional networks.

Supplementary material

Supplementary information – Additional figures and tables related to this work can be obtained from the webpage below.
References


