Computational Biology and Chemistry 33 (2009) 261–268

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

Plasticity of transcriptional machinery in bacteria is increased by the repertoire of regulatory families

Sarath Chandra Janga [a,*,1], Ernesto Pérez-Rueda [b,*,2]

[a] MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH, UK
[b] Departamento de Ingeniería Celular y Biocatálisis, Instituto de Biotecnologia. Universidad Nacional Autónoma de México, Cuernavaca, Morelos 62100, Mexico

ARTICLE INFO

Article history:
Received 30 September 2008
Received in revised form 16 June 2009
Accepted 17 June 2009

Keywords:
Transcription factor families
Regulatory network
Transcription machinery
Prokaryotes
Evolution

ABSTRACT

Escherichia coli K12 and *Bacillus subtilis* 168 are two of the best characterized bacterial organisms with a long history in molecular biology for understanding various mechanisms in prokaryotic species. However, at the level of transcriptional regulation little is known on a comparative scale. Here we address the question of the degree to which transcription factors (TFs) and their evolutionary families are shared between them. We found that 59 proteins and 28 families are shared between these two bacteria, whereas different subsets were lineage specific. We demonstrate that majority of the common families expand in a lineage-specific manner. More specifically, we found that AraC, ColD, Ebp, LuxR and LysR families are over-represented in *E. coli*, while ArsR, AsnC, MarR, MerR and TetR families have significantly expanded in *B. subtilis*. We introduce the notion of regulatory superfamilies based on an empirical number of functional categories regulated by them and show that these families are essentially different in the two bacteria. We further show that global regulators seem to be constrained to smaller regulatory families and generally originate from lineage-specific families. We find that although TF families may be conserved across genomes their functional roles might evolve in a lineage-specific manner and need not be conserved, indicating convergence to be an important phenomenon involved in the functional evolution of TFs of the same family. Although topologically the networks of transcriptional interactions among TF families are similar in both the genomes, we found that the players are different, suggesting different evolutionary origins for the transcriptional regulatory machinery in both bacteria. This study provides evidence from complete repertoires that not only novel families originate in different lineages but conserved TF families expand/contrast in a lineage-specific manner, and suggests that part of the global regulatory mechanisms might originate independently in different lineages.

1. Introduction

The genomes of the two model organisms, *Escherichia coli* K12 (Blattner et al., 1997) and *Bacillus subtilis* 168 (Kunst et al., 1997), contain a different proportion of Transcription Units (TU's) (Moreno-Hagelsieb and Collado-Vides, 2002), sigma factors and promoters (Salgado et al., 2006; Makita et al., 2004). Despite these basic differences, it has been possible to find some conserved and unique DNA-binding transcription factors (TFs) acting over their complete gene repertoires (Makita et al., 2004; Perez-Rueda et al., 2004). Such TFs have been related to a wide diversity of functions including catabolite repression, differentiation and cellular maintenance, among others. However, it is unclear how the collection of proteins performing similar functions (DNA-binding ability) could have evolved in these two organisms with different evolutionary history and ancestry (Hedges, 2002). Understanding the evolution of the transcriptional regulatory machinery across genomes would improve our knowledge about the evolutionary constraints that play a role in the formation of regulatory networks and would also help to decipher the design principles governing these networks across bacteria (Janga et al., 2009). Although some recent works have dealt with the evolution of the components and suggested duplication of genes as the main factor contributing to the formation of the Transcriptional Regulatory Network (TRN) (Madan Babu and Teichmann, 2003; Teichmann and Babu, 2004), there has not been comparative analysis of TFs and their families between genomes to understand the evolutionary constraints, functional aspects and design principles governing their formation. Despite the fact that there has been an increasing interest to identify and understand the regulatory repertoires of entire genomes using a variety of computational approaches (Perez-Rueda

* Corresponding authors.
E-mail addresses: sarath@mrc-lmb.cam.ac.uk (S.C. Janga), erueda@ibt.unam.mx (E. Pérez-Rueda).
[1] Tel.: +44 1223 402479; fax: +44 1223 213556.
[2] Tel.: +52 56 22 76 10; fax: +52 777 3 17 23 88.

1476-9271/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2009.06.004

et al., 2004; Brune et al., 2005; Moreno-Campuzano et al., 2006; Kummerfeld and Teichmann, 2006), there has not been genome scale comparative study reported so far to our knowledge, using representative genomes from distant lineages especially in the context of regulatory networks. Here we present the first comprehensive comparative analysis of the complete repertoires of TFs from two prokaryotic model organisms, *E. coli* K12 and *B. subtilis*.

In this work, we first identify and classify the repertoire of DNA-binding TFs of *E. coli* and *B. subtilis* into families using a previously reported approach applied to *E. coli* (Perez-Rueda and Collado-Vides, 2000). We then analyze the collection of TFs and their TF families at various levels to deduce thereof the common set of regulatory genes and families and to infer specific tendencies of TFs. Our analyses were based on the collection of TFs reported and collected from two different databases: RegulonDB (Salgado et al., 2006) for *E. coli K12* and DBTBS (Makita et al., 2004) for *B. subtilis*. Additional literature look up was performed, to retrieve a more complete dataset of TFs in these organisms. Here, we demonstrate that although *E. coli* and *B. subtilis* contain a similar proportion of DNA-binding TFs, the majority of the TF families have expanded and evolved independently. The regulatory networks based on the set of well-known TFs in both genomes suggest that the functions of genes regulated by similar families could be different. These findings open diverse opportunities to understand the complex regulatory systems in different bacteria, beyond Proteobacteria and Firmicutes.

2. Materials and Methods

2.1. Identification of TFs and Construction of TF Families in B. subtilis

In order to identify the repertoire of TFs in *B. subtilis*, we used a combination of information sources and bioinformatics tools as reported earlier (Moreno-Campuzano et al., 2006). Briefly, 237 TFs were identified by an exhaustive analysis of three sources, those TFs identified from DBTBS, a database devoted to the gene regulatory mechanisms in *B. subtilis* strain 168 (Makita et al., 2004), TFs identified by the search of family-specific Hidden Markov Models (HMMs) reported previously (Perez-Rueda et al., 2004) from *E. coli* TFs (*E*-value threshold $\geq$10-3), and those TFs identified with the library of HMMs from the Superfamily database (*E*-value $\geq$10-3) (Madera et al., 2004). This HMM library is based on the sequences of domains collected in the Structural Classification of Proteins (SCOP) database (Hubbard et al., 1997) and is thus applicable for a structural classification of proteins. In summary, the final dataset included those proteins identified by HMMs, Superfamily searches, and the repertoire (manually curated) of TFs described in DBTBS. These proteins were classified into families by using HMMs deposited in the PFAM DB (Bateman et al., 2000), and aligned by using the program *hmmalign* from HMMer. Our final collection included 90 families in *E. coli* and 51 families for *B. subtilis*. Additionally, their corresponding HMMs were used to scan a collection of 234 genomes, including bacterial, archaeal and eukaryotic species, in order to determine their evolutionary emergence in different lineages (see Supplementary Material for a complete list of genomes analyzed and the number of TFs identified across genomes).

2.2. Data of Regulatory Interactions

Transcriptional regulatory interactions of *E. coli* K12 were obtained from RegulonDB (Salgado et al., 2006), which contains experimental information extracted from literature, whereas the regulatory interactions of *B. subtilis* were retrieved from DBTBS (Makita et al., 2004). Those interactions from the datasets where a sigma factor is known to control the expression of a gene were excluded. Therefore, a total of 1816 regulatory interactions were considered for *E. coli* while 745 were included from the *B. subtilis* TRN.

2.3. Identification of Orthologs

Orthologs are defined as proteins in different species that evolved from a common ancestor by speciation (Fitch, 1970) and usually have the same function. Our working definition of orthology consisted of BLASTP reciprocal best hits, which is a widely accepted notion for identifying functional orthologs and homologous genes were identified with an *E*-value cutoff of 1e-6 as described elsewhere (Janga and Moreno-Hagelsieb, 2004).

3. Results and Discussion

3.1. Conserved TFs and TF Families Between E. coli and B. subtilis Genomes

Two proteins associated to common functions might be a consequence of common origin in different genomes (orthologous) or gene duplication within a genome after speciation (paralogous). Thus, we sought to determine the fraction of the total repertoire of TFs in *E. coli* and *B. subtilis* related by orthology and how it compares with genomic conservation. We found that 59 TFs from *E. coli* which correspond to around 20% of total TFs, had orthologs in *B. subtilis*, while around 29% of their total gene products are related by orthology which is statistically significant (see Supplementary Material), as has been previously observed about their conservation patterns using only a known subset of TFs in these genomes (Madan Babu et al., 2006; Lozada-Chavez et al., 2006). This finding suggests that TFs between the two genomes are 30% less conserved than other protein classes, indicating that TFs are likely lost to a greater extent at such phylogenetic distances (Lozada-Chavez et al., 2006). These observations give rise to several questions concerning the evolutionary and functional conservation of TFs between these bacterial genomes, so in order to have an insight into the commonalities and differences in the gene regulation between the prokaryotic species from the perspective of TFs, we used the complete repertoires of TFs in *E. coli* and *B. subtilis*. Based on diverse sequence and HMM searches, a total number of 303 *E. coli* TFs and 237 *B. subtilis* TFs were identified. These repertoires were also classified into families and compared to understand their evolutionary trends. Fig. 1 evidences the different proportions of TF families identified in the genomes. However, it can be noted that ArgR, BirA, DnaA, FrvR, LexA, PrpD and WrbA families show a very similar distribution in both the genomes. The similar proportion of these groups suggests the possibility of an early evolution of these families before the split of Proteobacteria and Firmicutes and no subsequent lineage-specific expansion or loss. A closer look at the functions of these families indicates that they are mostly involved in the synthesis of amino acids, replication and DNA repair mechanisms and metabolism of sugars. On the contrary AraC, ColD, DeoR, Ebp, IclR, LacI, LuxR, RpiR, YjhU_YdeW and YeiL families are dominant in *E. coli*, whereas ArsR, AsnC, GntR, Fur, MarR, MerR, ROK, TetR and OmpR can be seen to be dominant in *B. subtilis*. It is interesting to observe that AraC, ColD, Ebp, LuxR and LysR families are roughly double in proportion in *E. coli* than in *B. subtilis*, while ArsR, AsnC, MarR, MerR and TetR show a marked over-representation in *B. subtilis*. To test the significance of this observation and to determine if these distributions are in fact very different we performed a chi-square test, with the expected distribution in each genome calculated as the product of the total TFs from the common families and proportion of the TF family as seen in other genome. We observed a *P*-value $< 10^{-53}$ when the familial distribution in *B. subtilis* was considered as the observed