Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

A genomic-scale search for regulatory binding sites in the integration host factor regulon of *Escherichia coli* K12

M. Trindade dos Santos and Paulo Sérgio Rodrigues

Laboratório Nacional de Computação Científica - LNCC,
Av. Getúlio Vargas, 333, 25651-075 Petrópolis, RJ, Brasil
Corresponding author: M.T. Santos
E-mail: msantos@lncc.br

Genet. Mol. Res. 4 (4): 783-789 (2005)
Received February 24, 2005
Accepted November 23, 2005
Published December 27, 2005

ABSTRACT. We examined general aspects of the DNA-protein interaction between the integration host factor (IHF) global regulator and its regulatory binding sites in the *Escherichia coli* K12 genome. Two models were developed with distinct weight matrices for the regulatory binding sites recognized by IHF. Using these matrices we performed a genome scale scan and built a set of computationally predicted binding sites for each of the models. The sites found by the model associated with repetitive sequences had a higher score in the sequence to matrix alignment. They were also more rare than the other sites. The sites not associated with repeats rapidly tended to become undistinguishable from the background as statistical stringency was relaxed. We compared our results to the known sites documented in RegulonDB and found new members of the IHF Regulon. The two models exhibit clearly distinct affinity patterns (scores in the sequence to matrix alignments and in the number of regulatory sites), as we vary the stringency of the statistical confidence parameters. We suggest that these differences may play an important role in the dynamics of the network. We concluded that IHF may regulate two genes encoding ATP-dependent RNA helicases. This

 ©FUNPEC-RP www.funpecrp.com.br

interaction is not described in RegulonDB, even as a computational prediction. IHF may also regulate RNA modification processes.

Key words: IHF protein, Weight matrix, *Escherichia coli*

INTRODUCTION

Prokaryotes have the capability to elaborate genetic networking responses to adapt themselves to environmental changes. Transcriptional control of co-regulated genes is the main mechanism through which these responses are effected. However, to acquire a better understanding of gene network dynamics, there is a need to have a detailed description of all the elements acting on transcription-regulation control. These elements include promoters, transcription factor binding sites, terminator hairpins, and operons (Salgado et al., 2000). The topology of the DNA molecule itself (DNA bendability) has also been shown to play an important role in these processes (Stuart, 2000; Jauregui et al., 2003). A genomic-scale description of the regulatory network is also of fundamental importance to the analysis and interpretation of microarray data (Gutierres-Rios et al., 2003) and to the construction of a computational framework for modeling regulated metabolic networks. The role of the computational modeling of genetic and metabolic networks is to enable scientists to elaborate hypotheses and to infer biological behavior (Covert and Palsson, 2002). We examined general characteristics of regulatory binding site recognition by integration host factor (IHF), a global regulator in *Escherichia coli* (Lin and Lynch, 1996). One of the most complete resources of information on *E. coli* K12 regulatory elements and relations is the RegulonDB database (Salgado et al., 2004).

The first documented function of the IHF protein was its effect on *in vitro* integrative recombination of Lambda phage. Its function as a transcription factor had been firstly identified for genes related to transposition and plasmids. The IHF protein has also been found to be associated with regulation of the transcription of ribosomal genes, polymerases, other genes related to cell division, DNA replication and repair, flagellum genes, and a type III secretory system (Jauregui et al., 2003; Salgado et al., 2004). The regulation of such genes may also be indirect, i.e., IHF regulates the transcription of a gene that encodes another transcription factor, which regulates the transcription of other genes. As an example, we have the FIS protein, which is regulated by IHF, and is known to also regulate the transcription of ribosomal and polymerase genes (Jauregui et al., 2003). The RegulonDB Web site also provides information on the connectivity of the *E. coli* genetic network. This can lead to genetic circuits among the regulators (Thieffry et al., 1998). The IHF protein has some important specific properties, such as its association with topological changes in genomic DNA. The binding of IHF protein to its regulatory site involves bending of at least 140° in the DNA molecule. This bending and/or loop formation can bring elements close together (other transcription factors, regulatory binding sites and promoters as an example) that are far apart on the DNA molecule. DNA bendability is greater in the regulatory regions and has been shown to be conserved in these regions of bacterial genomes (Jauregui et al., 2003). Another feature is the high degree of synergistic interac-

tions; almost 30% of the known promoters regulated by IHF are also regulated by other transcription factors, such as CRP and FNR (Salgado et al., 2004). Besides the fact that IHF is a protein of fundamental importance for the regulatory network of *E. coli*, we chose to examine this protein because it is capable of recognizing two distinct patterns of regulatory binding sites, i.e., there are two consensus sequences and two frequency matrices as models for the binding sites recognized by this protein (Lin and Lynch, 1996). The sequences used to build one of the models are flanked by palindromic repetitive elements (REP), found more than a hundred times in *E. coli* intergenic sequences (Rudd, 1999). The role of intergenic repeats in bacterial genomes is not completely understood. We investigated the global aspects of the energy of the DNA-IHF-protein interaction, on a genomic scale in *E. coli*.

MATERIAL AND METHODS

We downloaded data on *E. coli* K12 strain MG1655 from genebank (http://www.ncbi.nlm.nih.gov/) (Blattner et al., 1997) and used the frequency matrices as models for the IHF DNA binding site sequences (Lin and Lynch, 1996). The frequency matrices are given by

$$A = \begin{vmatrix} 15 & 7 & 12 & 16 & 13 & 6 & 6 & 3 & 11 & 14 & 9 & 8 & 13 & 14 & 20 & 21 & 5 & 0 & 27 & 21 & 5 & 15 & 19 & 6 & 2 & 0 & 20 \\ 6 & 14 & 7 & 7 & 9 & 12 & 13 & 13 & 5 & 7 & 14 & 13 & 10 & 8 & 7 & 5 & 22 & 0 & 0 & 0 & 11 & 5 & 2 & 2 & 24 & 26 & 0 \\ 2 & 1 & 6 & 1 & 3 & 8 & 5 & 9 & 7 & 5 & 1 & 4 & 2 & 4 & 0 & 0 & 0 & 27 & 0 & 3 & 9 & 4 & 3 & 6 & 0 & 1 & 0 \\ 4 & 5 & 2 & 3 & 2 & 1 & 3 & 2 & 4 & 1 & 3 & 2 & 2 & 1 & 0 & 1 & 0 & 0 & 0 & 3 & 2 & 3 & 3 & 13 & 1 & 0 & 7 \end{vmatrix}$$

(Equation 1)

and

$$B = \begin{vmatrix} 26 & 2 & 26 & 27 & 21 & 25 & 3 & 0 & 9 & 0 & 1 & 3 & 26 & 26 & 28 & 2 & 0 & 0 & 28 & 28 & 0 & 27 & 8 & 27 & 0 & 0 & 0 \\ 2 & 2 & 0 & 1 & 1 & 2 & 14 & 6 & 5 & 24 & 1 & 1 & 1 & 0 & 0 & 26 & 28 & 0 & 0 & 0 & 25 & 0 & 16 & 0 & 28 & 28 & 0 \\ 0 & 24 & 2 & 0 & 0 & 0 & 5 & 22 & 3 & 3 & 1 & 24 & 1 & 0 & 0 & 0 & 0 & 28 & 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 6 & 1 & 6 & 0 & 11 & 1 & 25 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 3 & 1 & 1 & 1 & 0 & 0 & 28 \end{vmatrix}$$

(Equation 2)

which indicate the position of the binding sites used to build the matrix. The dimension of the matrices is (4 x W): four is for the letters of the alphabet and W is the length of the sites. The corresponding order for the letters in the lines is A, T, C, and G.

The IHF protein binds to the DNA as a dimer with non-identical units. The same protein is capable of recognizing these two clearly distinct classes of DNA binding sites. Regulatory sites used to construct Matrix B are flanked by REP (Rudd, 1999). Sites used to construct Matrix A had no particular flanking sequences. We wrote a C++ code that, given a frequency matrix and a DNA sequence of length L, opens a window of size W on the DNA sequence at position Pi (i ranging from 1 to (L - W +1)) and calculates the score. The frequency matrix is transformed into a weight matrix according to the following equation (Hertz and Stormo, 1999):

$$w_{i,j} = \frac{\left(n_{i,j} + p_i\right)/\left(N+1\right)}{p_i} \approx \ln \frac{f_{i,j}}{p_i} \qquad \text{(Equation 3)}$$

where $n_{i,j}$ is the number counted for base i at position j of the sites, N is the number of sites used to build the matrix, $f_{i,j} = n_{i,j}/N$ and p_i is the probability of finding base i at an arbitrary position of the genome ($p_i = 0.25$ for any base in our case).

For each sequence test in a window the score S is calculated as follows:

$$S = \sum_j \sum_i w_{ij} \delta_{il} \qquad \text{(Equation 4)}$$

where l is the letter of the alphabet in the test sequence at position j. We have

$$\delta_{il} = \begin{cases} 1 \text{ if } i = l \\ 0 \text{ if } i \neq l \end{cases} \qquad \text{(Equation 5)}$$

and thus only the letter of the matrix that is equal to the letter in the sequence test contributes to the calculation.

This score can be interpreted as the affinity of the protein for the regulatory DNA site modeled by the respective matrix. The code returns the score of the alignment between the substring within the window and the matrix, and also gives its coordinates on the complete genomic sequence. The window slides over the complete input sequence (Hertz and Stormo, 1999). Since the IHF binding site does not present a palindromic symmetry, we calculated the score for every position on both strands of the complete sequence of the *E. coli* K12 genome. The higher the score of a match, the higher the affinity of the sequence to the matrix (model). We then sorted the sequences by their score in descending order.

We expected to find functional regulatory binding sites in extragenic regions rather than in coding regions, and we used this to define one of the cutoff criteria for considering matches as reliable for further analysis; for each sequence in the list ordered by its score we determined if the coordinates of the center of the sliding window were on a coding or non-coding region of the genome (we considered as non-coding the first +50 5'-translated bases of each gene). A counter was used for sequences found in the non-coding regions and another counter for all sequences. This process was iterated until the number of sequences found in non-coding regions was half of the number of sequences read in the original list. The score of the last element of this iteration defines the cutoff value *a posteriori*. This criterium had previously been implemented in a similar way (Tan et al., 2001). As in most of the methods for searching for regulatory signals in DNA sequences, the main drawback here was the false-positives and false-negatives in the set of selected sequences. We also implemented independent criteria to study our data set and to ensure statistical significance. For a given matrix, we took the maximum score found in the alignments and considered as reliable the sequences with a score of at least 75% of this value. Data for the reliable matches were then parsed to a relational database (MySQL). The C++ code for scanning the sequences is available upon request from the authors.

RESULTS

We proceeded with our analysis in two directions. First, we examined global characteristics of IHF recognition and binding affinity to the putative DNA regulatory sites (Trindade dos

Santos M and Rodrigues PS, unpublished data). We made a search for distinct classes of binding sites associated with a single transcription factor, using information provided by the literature (Lin and Lynch, 1996; Blattner et al., 1997). By means of the frequency matrices (translated into weight matrices) A and B (Lin and Lynch, 1996), we constructed a set of statistically significant computationally predicted binding sites. Then, we examined the variation in site number and score as a function of statistical stringency parameters. The analysis was performed in a genome-wide context. We also found specific binding site sequences with a very high statistical confidence, indicating novel elements for the IHF Regulon in *E. coli*.

We took the highest score value, Smax, for each matrix and plotted the distribution (positions) over the complete genome of all the sequences with a score higher than this maximum multiplied by a constant. The distributions over the genome of the sequences having scores above (a) Score > Smin = Smax*0.75 and (b) Score > Smin = Smax*0.6 were plotted (Figure 1). The percentage of sequences counted outside coding regions was 97% for Matrix B, in both distributions, and 94% for Matrix A in distribution a and 74% in distribution b. Relaxing of stringency criteria from distribution a to distribution b produced an almost 4.5-fold increase (from 33 to 150) in the number of sequences found by Matrix A, whereas the matches by Matrix B remained almost the same (increasing from 40 to 47).

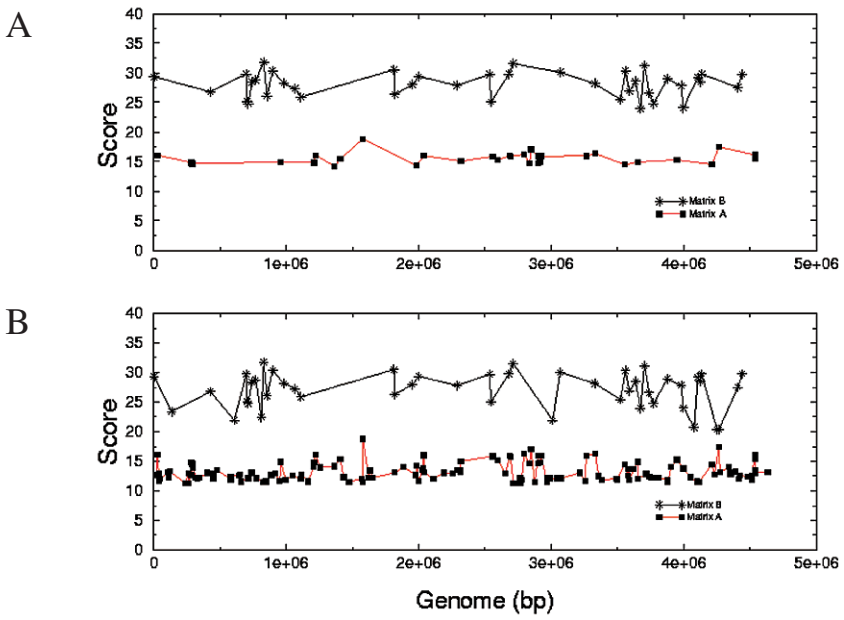

Figure 1. Positions over the complete genome of *Escherichia coli* of the highest scoring binding sites according to frequency matrices A and B (see Material and Methods). **A.** Binding sites with score above Smax*0.75. **B.** Binding sites with score above Smax*0.60. Smax is the maximum value found for the score in the genome. Scores are given in arbitrary units.

Sites recognized by Matrix B are less in number but they bound with higher energy (score). Sites recognized by Matrix A rapidly degenerated, increasing in number and binding with a low score, as the criterium was relaxed. We calculated the mean score and the standard deviation (SD) for the complete genome. Matrix A: we found mean score = -18.4, SD = 6.81,

and Matrix B mean score = -41.61 and SD = 9.61. In the worst case, the low score for Matrix A was 4.36 SD units from the mean, while for Matrix B the minimum distance was 6.44 SD units (Figure 1).

If we implement a criterium such as described in Tan et al., 2001, taking as cutoff value the percentage of sites in non-coding regions, the results are more expressive; using the cutoff value for the score corresponding to 60% of the sequences located in extragenic regions, we found 198 putative binding sites recognized by Model B, regulating the transcription of 185 open reading frames, and 720 binding sites recognized by Model A, regulating the transcription of 575 open reading frames. This is a clear indication that the IHF Regulon is much larger than we originally expected.

We also compared our results to the known sites in RegulonDB (Salgado et al., 2004), using the same criteria. Among the 55 transcription units for which there is a promoter known to be regulated by the IHF, we recognized 30 (55%). Surprisingly, most of the known sequences (77%) were best recognized by the low-scoring Matrix A model.

When we examined specific sequences, we found that the two highest-scoring matches were located at the promoter regions of genes rhlE (starting the match at -119 bp, a sequence given by ACAAAAGCATGCAAATTCAATATATTG in the direction of transcription) and srmB (starting the match at -119 bp, a sequence given by ACAAGATCGTGCAAATTCAATATATTG in the direction of transcription). The +1 position is the first base of each gene. Both genes are annotated as ATP-dependent RNA helicases, suggesting that IHF also regulates these genes related to RNA modification processes. These two regulatory sites were not described in RegulonDB and were recognized by Matrix B.

Finally, we calculated the information content for the two matrices, A and B, and for the matrix presented without separation by the models. The information content (Hertz and Stormo, 1999) is a measure of the capability of the model to distinguish between a functional site and an arbitrary sequence from the background. We found that the least information content is generated when the models are not used to separate the data. We found I(A) = 352.10 for Matrix A, I(B) = 777.81 for Matrix B and I(R) = 308.59 (given in arbitrary units). The value for I(R) was calculated from the matrix given by the RegulonDB Web site. The difference in the information content found for I(A) and I(B) also reflects the difference in the scores shown in Figure 1.

DISCUSSION

The two models for DNA binding recognition, for the IHF protein presented in Lin and Lynch, 1996, gave different binding energies and degeneracy. Sites recognized by Matrix A are much more numerous, but they had a low score. On the other hand, Matrix B models a set of sequences with a high affinity for the protein (high score for the sequence to matrix alignment). The DNA bending and loop structures associated with the IHF binding can induce an increase in the local density of the other regulatory proteins. These topological modifications of the DNA molecules are so dramatic that the bending itself may be functional. In principle, the choice of IHF for binding to a site according to one model or the other may not only depend on thermodynamic constraints (Jauregui et al., 2003) (the IHF protein does not require interaction with any co-factors to recognize its binding sites) and may be directed by the regulatory network interactions. The score in the sequence to matrix alignment is not an indication of whether the protein binds or not to a given site, but it is related to the time that such binding lasts.

This particularity of binding with different affinities in the different models may play an important role in the dynamics of the regulatory network. For these reasons, we believe that this is one of the *E. coli* proteins responsible for coordinating complex responses to complex environmental stimuli (Kitano, 2002).

We separated the search into two models. This made it possible to identify (with high statistical significance) novel (putative) binding sites for the IHF regulon, indicating for example that this protein may also regulate RNA metabolism processes through control of the transcription of the genes rhlE and srmB (both recognized by Model B). Since one of the models indicates sequences associated with the intergenic repeats in bacteria, we can question the role of these repeats in gene regulation. Is there any correlation between the occurrence of intergenic repeats and other transcription factor binding sites? Do repetitive sequences play any specific role in the dynamics of genetic and metabolic networks? These are relevant questions to be investigated in future work.

ACKNOWLEDGMENTS

We thank Laboratório Nacional de Computação Científica - LNCC/MCT for financial support.

REFERENCES

Blattner FR, Plunkett G, Bloch CA, Perna NT et al. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science* 277: 1453-1474.

Covert MW and Palsson BO (2002). Transcriptional regulation in constraints-based metabolic models of *Escherichia coli*. *J. Biol. Chem.* 277: 28058-28064.

Gutierres-Rios RM, Rosenblueth DA, Loza JA, Huerta AM et al. (2003). Regulatory network of *Escherichia coli*: consistency between literature knowledge and microarray profiles. *Genome Res.* 13: 2435-2443.

Hertz GZ and Stormo GD (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* 15: 563-577.

Jauregui R, Abreu-Goodger C, Moreno-Hagelsieb G, Collado-Vides J et al. (2003). Conservation of DNA curvature signals in regulatory regions of prokaryotic genes. *Nucleic Acids Res.* 31: 6770-6777.

Kitano H (2002). Systems biology: a brief overview. *Science* 295: 1662-1664.

Lin ECC and Lynch AS (1996). Regulation of gene expression in *Escherichia coli*. R.G. Landes Company and Chapman & Hall, R.G. Landes Company, Georgetown, TX, USA.

Rudd KE (1999). Novel intergenic repeats of *Escherichia coli* K-12. *Res. Microbiol.* 150: 653-664.

Salgado H, Moreno-Hagelsieb G, Smith TF and Collado-Vides J (2000). Operons in *Escherichia coli*: genomic analyses and predictions. *PNAS* 97: 6652-6657.

Salgado H, Gama-Castro S, Martinez-Antonio A, Diaz-Peredo E et al. (2004). RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acids Res.* 32: 303-306.

Stuart MA, Long AD, Ito ET, Tolleri L et al. (2000). Global gene expression profiling in *Escherichia coli* K12: The effects of integration host factor. *J. Biol. Chem.* 275: 29672-29684.

Tan K, Moreno-Hagelsieb G, Collado-Vides J and Stormo GD (2001). A comparative genomics approach to prediction of new members of regulons. *Genome Res.* 11: 566-584.

Thieffry D, Huerta AM, Perez-Rueda E and Collado-Vides J (1998). From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli*. *Bioessays* 20: 433-440.