Biochimie 83 (2001) 201–212

S0300908400012256/FLA

Genome organisation and chromatin structure in *Escherichia coli*

David Ussery*, Thomas Schou Larsen, K. Trevor Wilkes**, Carsten Friis, Peder Worning, Anders Krogh, Søren Brunak

Center for Biological Sequence Analysis, Department of Biotechnology, Building 208, The Technical University of Denmark, 2800 Kgs. Lyngby, Denmark

(Received 30 November 2000; accepted 15 December 2000)

Abstract — We have analysed the complete sequence of the *Escherichia coli* K12 isolate MG1655 genome for chromatin-associated protein binding sites, and compared the predicted location of predicted sites with experimental expression data from 'DNA chip' experiments. Of the dozen proteins associated with chromatin in *E. coli*, only three have been shown to have significant binding preferences: integration host factor (IHF) has the strongest binding site preference, and FIS sites show a weak consensus, and there is no clear consensus site for binding of the H-NS protein. Using hidden Markov models (HMMs), we predict the location of 608 IHF sites, scattered throughout the genome. A subset of the IHF sites associated with repeats tends to be clustered around the origin of replication. We estimate there could be roughly 6000 FIS sites in *E. coli*, and the sites tend to be localised in two regions flanking the replication termini. We also show that the regions upstream of genes regulated by H-NS are more curved and have a higher AT content than regions upstream of other genes. These regions in general would also be localised near the replication terminus. © 2001 Société française de biochimie et biologie moléculaire / Éditions scientifiques et médicales Elsevier SAS

IHF / H-NS / FIS / bacterial chromatin / repetitive DNA / DNA curvature / *Escherichia coli*

1. Introduction

1.1. Chromatin proteins in E. coli

There are 12 chromatin proteins found in *E. coli*, although only five exist in abundant concentrations (e.g., 20 000 copies per cell or more); historically, four have been well characterised in *E. coli*: HU, IHF, H-NS, and FIS [16]. In stationary phase, HU and FIS are replaced by the Dps protein [24]. These proteins play a major role in organising the *E. coli* nucleoid structure in the roughly thousand-fold compacted chromosomal DNA. As in eukaryotes, chromatin structure plays an important role in gene expression, and the interplay between the chromatin-associated proteins can be involved in the regulation of specific genes [23]. Recent experiments have shown that whilst many of the chromatin-associated proteins in *E. coli* are localised throughout the genome, some (including FIS) are localised to only a few regions within the genome [2].

* *Correspondence and reprints.*
E-mail address: dave@cbs.dtu.dk (D. Ussery).
** Present address: Department of Biology, Roanoke College, Salem, Virginia 24153-3794, USA.
Abbreviations: IHF, integration host factor; FIS, factor for inversion stimulation; H-NS, high molecular mass native subunit.

1.2. IHF binding and stabilisation of DNA looping

IHF and HU are members of the histone-like protein family; IHF has been found to bind to a degenerate consensus site: YAACTTNTTGATTTW [19], whilst HU shows essentially no sequence preference. Both HU and IHF are heterodimers, and although the binding site for IHF is asymmetrical, the orientation of the IHF binding site relative to the promoter can be reversed with no apparent loss of activity in the promoter region of bacteriophage Mu [51]. The crystal structure of an IHF dimer bound to DNA [36] reveals a minor groove docking ribbon motif [6], which induces a nearly 180 degree turn over about 30 bp of DNA, upon the binding of one IHF dimer, as shown in *figure 1A*. The binding of IHF is greatly influenced by structural features within the DNA helix, rather than by the nucleotide sequence alone [46]. The structure of the HU protein is shown in *figure 1B*. Although both HU and IHF share similar DNA binding folds, the HU protein shows little sequence preference, and does not wrap DNA around itself.

IHF got its name (integration host factor) based on its ability to facilitate the integration of lambda phage into *E. coli*; it is now known that this is achieved through specific looping of the DNA [13]. The exact location of the IHF binding site with respect to the lambda attachment site is important; recombination is reduced when the binding site is moved only one bp to either side [30]. Obviously, the role of IHF in *E. coli* is more than merely facilitating the

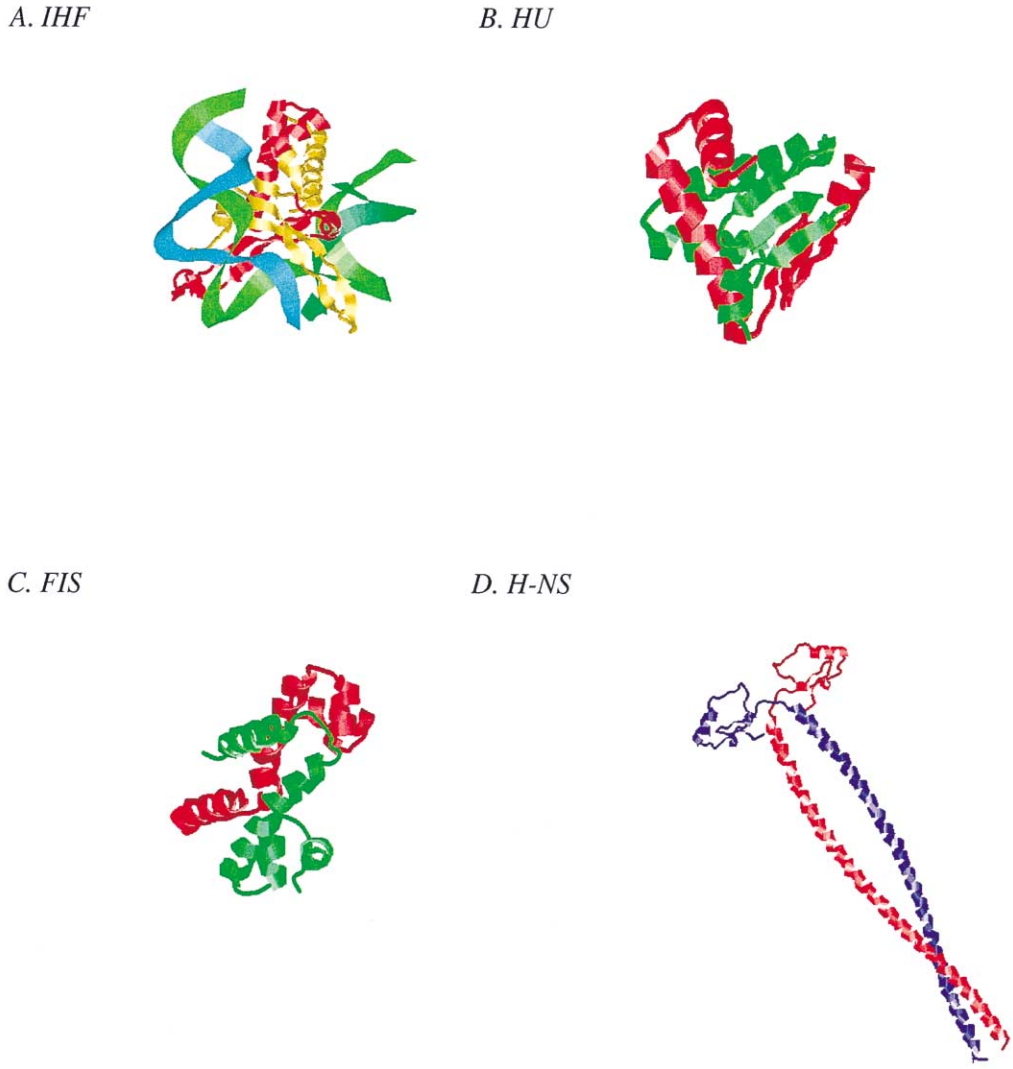

Figure 1. A. Structure of the IHF protein dimer, with DNA wrapped around it [6]. **B.** Structure of the HU protein from *Thermotoga maritima* (see http://www.ncbi.nlm.nih.gov/Structure/mmdb/mmdbsrv.cgi?form=6\&db=t\&Dopt=s\&uid=12816. **C.** Structure of FIS protein [56]. **D.** 3-D model for the *E. coli* H-NS protein. The structure for the N-terminal domain has been solved using 1-H NMR, but structure of the remaining two-thirds of the protein remains undetermined. Shown here is a model of the 3-D conformation, made using 'Copenhagen Models' at CBS (http://www.cbs.dtu.dk/services/CPHmodels/).

integration of phage lambda; IHF can modulate the transcriptional activity of promoters by influencing the looping of upstream DNA [18]. The number of IHF binding sites in the *E. coli* genome is presently unknown, but many of the documented IHF binding sites are localised within repetitive extragenic palindromic (REP) sequences [3] and have been named 'RIB' for reiterative IHF BIME (bacterial interspersed mosaic elements) [8, 31]. It is estimated that there are roughly 70 to 100 copies of the 'RIB' repeat in the *E. coli* genome.

1.3. FIS regulation of highly expressed genes

Binding of FIS protein to upstream regions has been found to enhance the transcription levels of highly expressed genes, such as rRNAs [1]. However, FIS can bind DNA in a non-specific manner as well as at specific sites, making determination of a clear consensus site difficult [5]. The FIS protein contains a helix-turn-helix motif [27], which binds DNA in the major groove; the structure of a FIS dimer is shown in *figure 1C*. FIS can also stabilise

DNA looping to enhance transcription [47]. Although FIS is considered a 'global regulator' of gene expression, there are only a handful of protein-encoding genes known to be activated by FIS. A recent study using 2-D gels found that only about 10 proteins were overexpressed due to FIS, in rich versus minimal growth media [11]. The amount of FIS protein varies with the growth phase, from less than 100 copies in stationary phase to over 500 000 copies per cell in log phase [4]. Estimates for the number of FIS binding sites in the *E. coli* genome range from a mere six [45] to more than 68 000 [22].

1.4. H-NS, chromatin structure, and global regulation of gene expression

H-NS is involved in the condensation of the bacterial chromosome, and can repress the transcription of several genes, presumably through alteration of chromatin structure [48]. A 3-D model of H-NS protein is shown in *figure 1D*; note that to date the structure for only the C-terminal third of the protein has been solved [39]. The H-NS protein is organised into two domains: an oligomerisation domain at the N-terminus, and a DNA binding domain that is roughly the C-terminus third of the protein [15]. Based on the sequence, it is likely that the H-NS 'oligomerisation domain' consists of a coiled-coil motif. Although H-NS was originally thought to exist primarily as a dimer, the oligomeric status of H-NS in the cell is presently thought to consist of tetramers and larger oligomeric combinations [41]. H-NS plays a major role in condensation [12, 43], and it is likely the reason many genes are transcriptionally repressed by H-NS is due to this compaction.

H-NS affects the expression of a set of specific genes, but an H-NS consensus binding site has not been found. It is likely that H-NS recognises specific DNA structures, rather than specific sequences [26]. In spite of the lack of a clear DNA binding consensus, mutants in the *hns* gene result in a variety of different phenotypes [53], and H-NS plays a major role as a global regulator in phosphate-starved cells [21]. There are three isoforms of H-NS [42], one of which is found at about four-fold higher concentrations in phosphate-starved cells [52]. There are roughly 20 000 copies of the H-NS monomer per cell [50].

1.5. DNA curvature

It has been known for some time that regions upstream of actively transcribed genes are often curved [34, 44]. Synthetic DNA curves inserted upstream of genes were found to activate transcription in vitro and also in vivo [9], although H-NS can bind to certain curves in upstream regions and thereby repress transcription [50].

Often the DNA curvature, both measured and calculated, is related to the phasing of short oligo A-tracts. This type of curvature is quite sensitive to temperature; whilst stable at temperatures below 30 °C, the curvature is greatly reduced at temperatures above 40 °C [10]. However, there are other sequence motifs which give rise to curvature which are not temperature-dependent [49]. H-NS might bind preferentially to AT-rich curves at lower temperatures, and not bind at the higher temperatures, enhancing the effects of the structural changes in the DNA. Thus it is possible that the temperature of the environment can act as a switch to modulate the level of expression of the gene. It is interesting to note along these lines that many of the genes regulated by H-NS are environmentally sensitive, including virulence genes [16]. Recently, a virulence plasmid in *Yersinia* has been found to be enriched in curves which melt at 37 °C, which is consistent with this form of regulation [37].

When one looks at the entire *E. coli* chromosome, in terms of DNA curvature, there is a clear tendency for more curved regions to be localised near the replication terminus [33]. The same trend is also seen for the *Bacillus subtilus* chromosome. Since it is known that curved regions are localised at the apical ends of supercoils, this would imply that a supercoiled *E. coli* chromosome would have the tips of plectonemic supercoils localised preferentially near the replication terminus.

1.6. Prediction of chromatin-associated protein binding sites

Of the major chromatin-associated proteins, only two (FIS and IHF) bind to specific sequences; H-NS shows a weak sequence/structure preference, and HU binds all DNA sequences with roughly equal affinity. Documented FIS and IHF binding sites are often found in clusters of three sites, near or within promoter regions. If one assumes that all *E. coli* genes have three sites, this would put an upper limit in the range of about 10 000 FIS and IHF sites. However, since it is likely that many genes are not regulated by IHF and FIS, the number could be much lower than this. All of the major chromatin-associated proteins exist at levels of at least 20 000 copies per cell; it is thus reasonable to expect to find many (perhaps a few thousand) potential binding sites throughout the chromosome.

2. Materials and methods

2.1. Hidden Markov models for predicting IHF and FIS binding

In order to capture characteristics of DNA sequences to which FIS or IHF are known to bind, we have chosen to use hidden Markov models (HMMs) because they can be estimated from unaligned sequences, unlike weight matrices and many other methods. Another advantage, when compared to techniques such as neural networks, is the ease of relating trained model parameters to sequence

information; for instance, it is possible to directly read off any consensus signals found and to get a good idea of the information present in these signals. For introductions to HMMs we recommend references [17, 28].

The central idea of an HMM is to embed the statistics of a motif in a set of states with transitions between them. Each HMM state has a specific probability distribution over the four nucleotides and hence one may say that it 'emits' nucleotides according to specific emission probabilities. There is a state for each position in the motif and the emission probabilities essentially end up being equal to the nucleotide frequencies at these positions. Hence, an HMM may be viewed either as a generative model which 'emits' nucleotides according to specific statistics or as a scoring model which may be used to answer questions such as: "To what extent is a given sequence compatible/similar to the sequences used to train the HMM?". These two HMM interpretations are equally valid and the choice between them depends on the application in question.

In our case, there is also a state modelling the sequence background as a first order Markov chain estimated from the entire *E. coli* genome. The emission probabilities are estimated by an iterative procedure known as the Baum-Welch algorithm by using a training set of unaligned sequences of experimentally verified binding sites. Starting from an arbitrary model the estimation procedure can be thought of as optimising (in a maximum likelihood sense) both the alignment of the sequences and the probability parameters of the model, so in the end it will have found the most significant motif in the training sequences. The parameters of the state modelling the background are fixed at all times and not changed by the estimation procedure. The transition from the background state to the first state of the binding site model has a fixed probability. The exact value of the transition probability from the background state is unimportant, because it just scales the posterior probability of a binding site, and thus only changes the threshold used to discriminate between binding sites and non-binding sites.

For predicting sites in the genome we calculate, for each nucleotide, the posterior probability that it is the middle of a binding site. During training we actually use two copies of the background state, so a path through the model always starts in the first background state, passes through the states modelling the binding site, and end up in the second background state. This forces the HMM to find one and only one binding site in each training sequence, and in this way the statistics of FIS or IHF regions were embedded in the HMM during training. The trained model can then be modified to allow for multiple sites in a sequence by merging the two background states and it can then be used to scan an entire genome.

A leave-one-out cross-validation was employed to measure the performance of the trained model, i.e., the HMM was trained on all sequences except one, which was then used to test the trained model (with the architecture slightly modified in order allow no or multiple occurrences of binding sites). The number N of test sequences scoring above some threshold could then be used to estimate the percentage of false negatives given by $\%FN = 100\,\frac{M-N}{M}$, where M is the total number of training sequences. Because of the cross-validation $\%FN$ is a reasonable estimate of the percentage of the actual sites in the *E. coli* genome that will be found by the HMM. By setting a very low threshold one could obviously achieve $\%FN = 0$, but the model would then overpredict, yielding too many false positives. Finding the optimal threshold is non-trivial and will be discussed below. The sensitivity of the model estimated from the cross-validation is simply $100 - \%FN$.

To get a rough idea of the false positive rate, the number of sites found using various thresholds were compared to those found in random sequences of the same length as the genome. These random sequences were generated from a first order Markov chain identical to the one used in the background state of the model. It should be emphasized that the number of sites found in random sequences is a rather crude estimate of the false positive extent, since true binding sites could actually be abundant in these sequences. This is especially true if the motif sought for has a weak consensus sequence.

2.2. The data sets

By searching the literature we found 78 training sequences (with lengths of 21 to 35 bp) of FIS binding sites and 67 sequences (with lengths of 34 to 48 bp) of IHF binding sites. The strand chosen for a binding site is the same as the strand of the down-stream gene, which the site is supposed to regulate according to the paper in which it is found. The sequences and references can be found in the supplemental section on our web page (http://www.cbs.dtu.dk/services/GenomeAtlas/chromatin). In order to check for strand asymmetry the following procedure was employed. A model with two identical submodels was trained on all the sequences and their reverse complements. If there was a clear split such that a sequence and its reverse complement always used two different submodels, the site was assumed to be strand asymmetric and for all the sequences we chose the strand matching the first submodel for the rest of the analysis.

Using the Viterbi algorithm [17] the most probable path of a sequence through the model was found. This path yields an alignment of the sequence to the model, and thus a multiple alignment of all the sequences in the training set. Such alignments were the basis for the construction of the sequence logos [38].

To study the location of IHF and FIS sites in promoter regions we extracted all the σ^{70} promoters annotated in the *E. coli* Genbank file with 1000 basepairs upstream of these sites. There was a total of 3929 such sequences.

2.3. GenomeAtlas plots

All results are reported for the *E. coli* strain K-12 MG1655 genome (GenBank accession number U00096 [7]). We used the 'GenomeAtlas' plots to visualise IHF and FIS binding sites, as well as DNA structural parameters, throughout the genome [25, 33]. Specific locations of the predicted web sites will be made available for inspection. In order to visualise the entire 4.6 million bp *E. coli* genome on one page, it was necessary to smooth all data over a 5000 bp window. mRNA concentrations were kindly provided by Garry Miyada (Affymetrix Inc., Santa Clara, California, USA). Protein concentrations were downloaded from the following web page: http://arep.med.harvard.edu/labgc/proteom.html.

3. Results and discussion

3.1. Genome organisation in E. coli

Figure 2 shows an overview of the *E. coli* genome. There are two sets of three circles. The outer set of three circles are arranged in the order of DNA coding regions (genes are labelled red or blue, depending on the orientation), mRNA expression levels (green circle; this is from data for cells grown in minimal media), and protein concentration levels (in blue). There are 4397 annotated genes in the *E. coli* genome, of which only 2005 are expressed at detectable levels in cells grown in minimal media as shown in the green circle; and only 233 proteins have been found to exist at 'abundant' concentrations (e.g., greater than about 100 molecules per cell), based on two-dimensional gel analysis [29].

The inner set of three circles corresponds to the predicted sites for IHF, IHF-RIB, and FIS proteins, respectively. The intensity of the colour reflects the predicted strength of binding. At level of the whole genome, the predicted binding sites are more or less distributed evenly throughout the genome. The next sections will discuss the predicted protein binding sites in more detail.

3.2. IHF

Using the dual-strand model described in the *Section 2.2.*, we found that the IHF binding site is strand asymmetric, and 32 of the 67 sequences in the IHF data set were shifted to the complimentary strand. After testing a few different model lengths, we found the best performing HMM to have 25 IHF motif states. An alignment of the sequences using that model produced the logo plot in *figure 3*. The consensus for the IHF binding site from Englehorn et al. [19] is shown below the logo, for comparison. Note that the central T is conserved in all 67 sequences from the training set of experimentally mapped IHF sites, and that there is considerable variation in the conservation of other nucleotides within the binding site.

We do not have a large set of negative sites, i.e., sites where IHF is known not to bind, and therefore it is very difficult to directly assess the specificity of the model. However, it can be done indirectly. In *figure 4* the number of IHF sites found in the genome (both strands) is shown versus the sensitivity to the training sequences found in cross-validation. Each point in *figure 4* corresponds to a specific threshold. The crosses are from the *E. coli* genome and the triangles are from a 4640 kb sequence generated from background statistics as described above. One sees that both curves have dramatic bends around a sensitivity of 0.72 indicating that this is the sensitivity level at which false positives start to dominate. Hence, we chose the threshold corresponding to this sensitivity level for the rest of the analysis. The percentage false negatives (IHF sites missed in the genome) at this threshold is then estimated to be 28%. Even at this level, the model does predict IHF sites in the random sequence. We believe that some fraction of those are actually functional IHF binding sites, although at present we cannot give a good estimate of the exact fraction.

Using this threshold, 307 sites were found on the direct *E. coli* strand and 315 on the reverse with 14 sites falling within 25 base pairs of each other on both strands yielding a total of 608 ((307 + 315) – 14) distinct sites found in the *E. coli* genome. Based on the extrapolation of the roughly linear curve for the true positives in *figure 4* we estimate that the actual number of IHF sites in *E. coli* is around 1000. However, this estimate may in fact be dependent on our specific training set (Richard Deonier, personal communication). Resolving this question would be an obvious line of future enquiry.

The density of predicted genomic IHF sites are shown in the violet ring of *figure 2*, and the next ring (red) shows IHF-RIB sites (see below). The predicted IHF binding sites are scattered throughout the chromosome. The IHF sites are expected to be concentrated immediately upstream of transcription starts and *figure 5* confirms this by showing a peak in the number of found IHF sites within 100 bp of the transcription start sites. *Figure 6* shows the separations of the sites found in the *E. coli* genome compared with a random sequence of genome length. Evidently, the real IHF sites are more clustered than one would expect by chance.

3.2.1. IHF-REP sites

For certain types of IHF sites the predictions can be significantly improved; 42 of the original 67 IHF sequences were from the *E. coli* genome and of these 42, 28 have highly conserved flanking regions known as REP sequences [31]. By adding flank (50 bp to each end) to the training sequences and adding 2×11 states to the model (making a total of 57 states), we improved the model's discrimination ability substantially; of course, at the cost

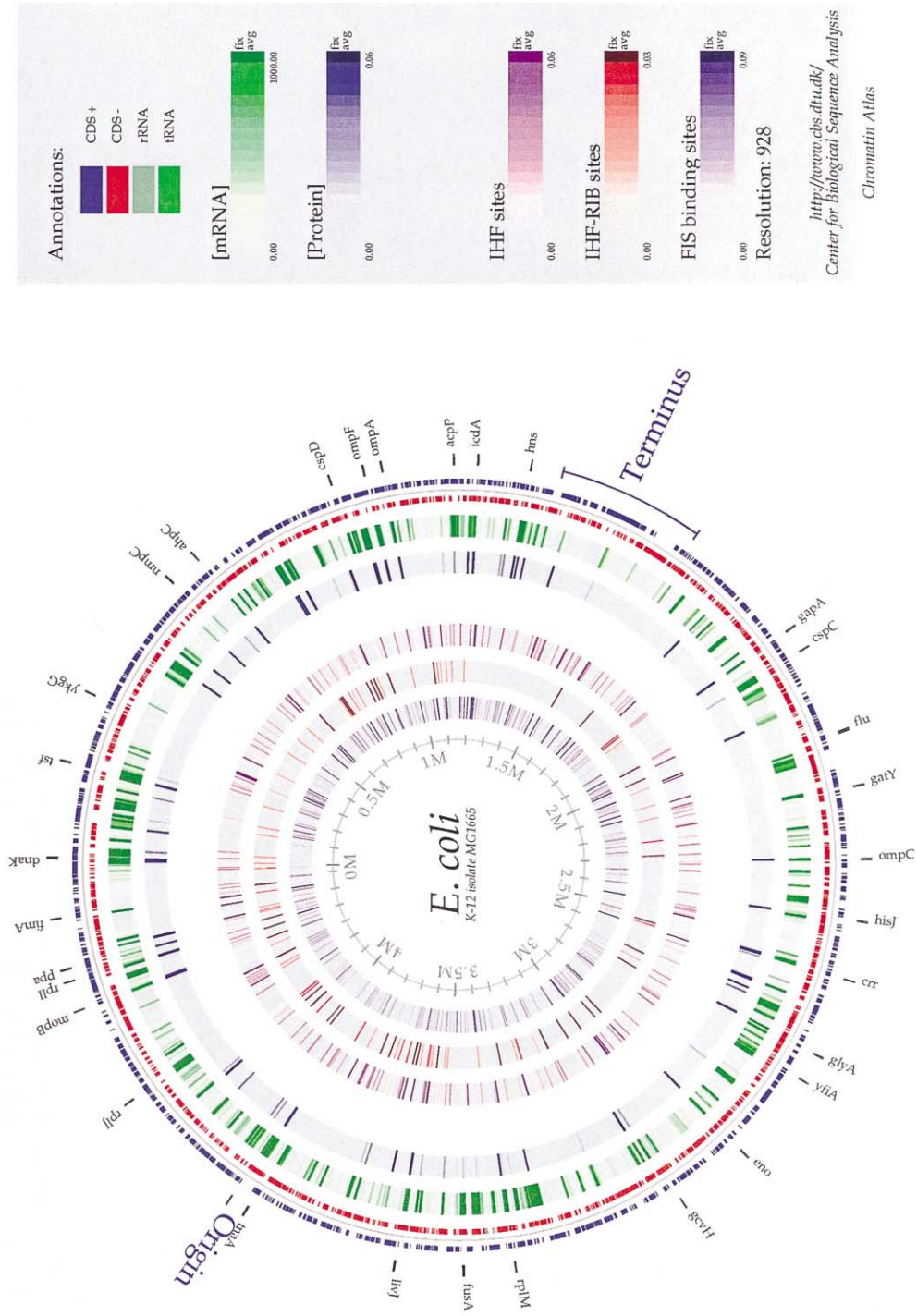

Figure 2. A. GenomeAtlas plot of the complete *E. coli* genome. The outer circle represents the genes (genes in the 'forward direction' are red, genes in the 'reverse direction' are blue). The next circle (green) is the mRNA concentration, based on levels of fluorescence from an Affymetrix 'DNA chip' experiment (the intensity of the colour reflects the concentration of the mRNA). The third circle is the concentration of 233 abundant proteins (again, the intensity of the colour is reflective of the relative concentration of the protein). The inner three proteins are the chromatin protein predicted binding sites, as discussed in the text.

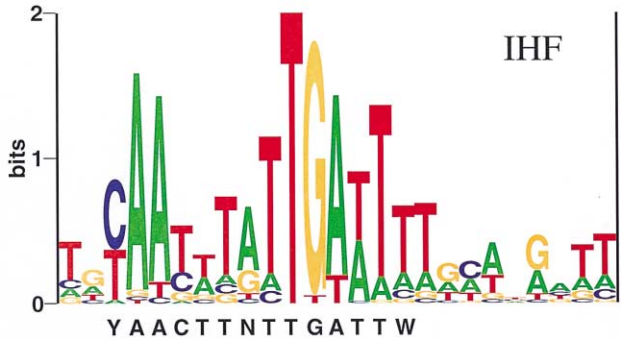

Figure 3. Logo plot for IHF binding sites in *E. coli*, with the consensus from Engelhorn et al. [19] underneath, for comparison. (Abbreviations for the nucleotides are: N = A,T,G, or C; Y = C or T; W = A or T). The logo shows information content in bits for each position in the alignment as the total height of letters. The relative height of the letters equals their frequencies [37].

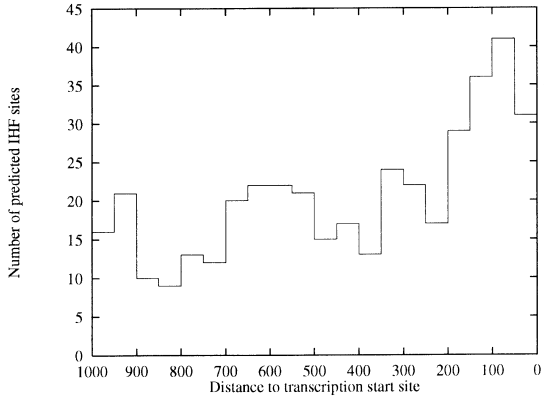

Figure 5. The spatial distribution of IHF binding sites 1000 bases upstream of all transcription start sites in the *E. coli* genome.

of only finding IHF sites with a REP-like flank. Scanning the *E. coli* genome with this model we found 88 hits on the direct strand and 87 on the reverse with 45 of these falling within 57 base pairs of each other indicating a total of 130 ((88 + 87) – 45) separate REP-like flanked IHF sequences in the genome. In this case only three sequences were found in a random sequence of genome length so we can be fairly certain that the sites found are real. We are also reasonably sure that all REP-like IHF sequences are found since the highly conserved flank made it possible to achieve $\%FN \approx 0$. The density of the found sites are shown in the red ring of *figure 2*. Note the high IHF-RIB density around the genome origin of replication, which is in agreement with experimental findings [35]. 63 of the sites found on the direct strand and 62 of those on the reverse strand were within 57 base pairs of the 333 REP sequences reported in the *E. coli* GenBank file [7]. 44 of these 125 sites overlap by 57 base pairs, which implies that 81 of the found 130 REP-like flanked sequences coincided with known REP sequences. The remaining 49 REP-like flanked sequences could be flanked by something not entirely like REP sequences, or it is possible that

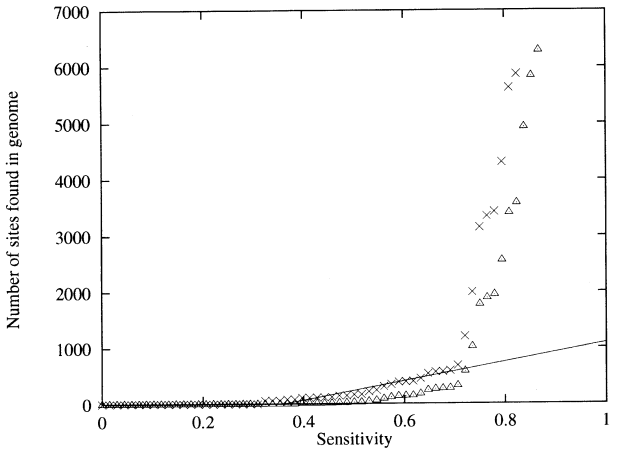

Figure 4. Estimation of the number of IHF sites in the complete *E. coli* genome as a function of the sensitivity as estimated in the cross-validation (crosses). The triangles show the number of predicted sites in a random sequence. The line is a crude extrapolation of the line believed to correspond to the correct predictions.

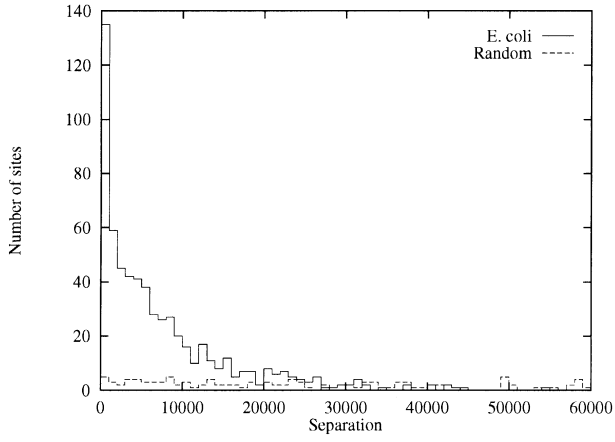

Figure 6. The distribution of separations (measures in base pairs) between IHF sites in the *E. coli* genome (full line) and, for comparison, in a random sequence of genome length (dashed line).

Table I. Prediction of IHF binding sites in the *E. coli* genome.

	Sites found	Estimated total
IHF	608	1 100
IHF-REP	130	130

not all REP sequences are currently annotated in the GenBank file *(table I)*.

3.3. H-NS regulated genes contain AT rich upstream regions

Previous experiments have indicated a relationship between DNA curvature and H-NS binding [32, 55, 57]. To investigate this, a dataset was created containing 500 bp regions upstream of translation start from 39 genes known to be regulated by H-NS. Using the CURVATURE programme [40] the average intrinsic curvature for the entire dataset was calculated and found to be around 10% higher than the average of upstream regions for all genes in *E. coli*. To evaluate the statistical significance of this observation, the average curvature was calculated for 100 000 datasets, each generated by randomly selecting 39 upstream regions from the entire genome, as shown in the upper panel in *figure 7*. Since there is a known correlation between curvature and base composition [14], a similar figure was created using the same procedure used for curvature except from AT content being calculated (*figure 7*, lower panel). The deviation from the upstream average is much larger for AT content than for curvature; at present it is unclear whether H-NS binding sites might be curved because H-NS binds to AT rich areas (which are known to be more curved), or the other way around. Efforts at finding a more specific consensus for H-NS binding, using the ANN-SPEC programme [54] were inconclusive.

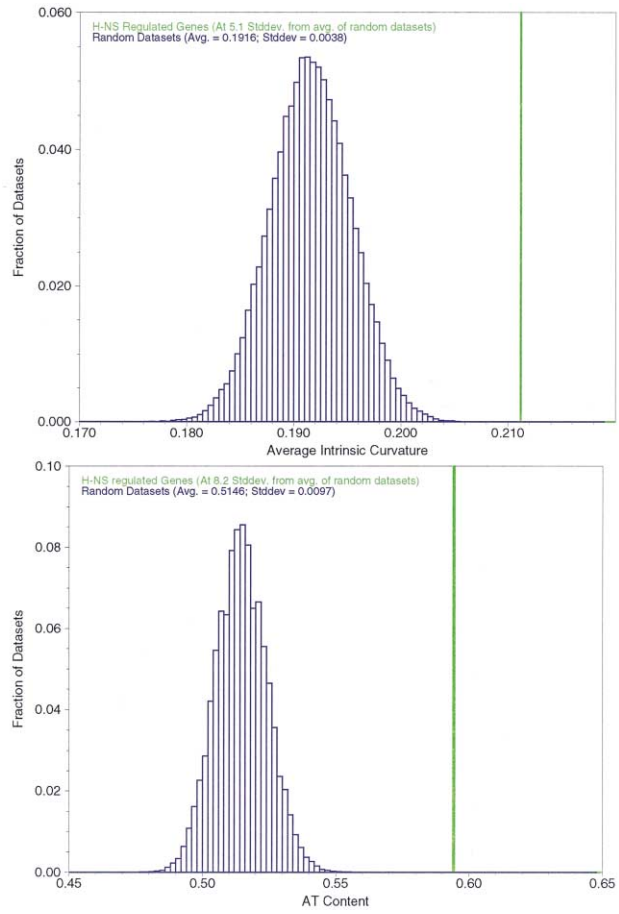

Figure 7. Comparison of the average curvature and AT content for upstream regions from H-NS regulated genes against datasets created by randomly selecting upstream regions from the entire genomes. For the latter, assuming a normal distribution, the average and standard deviation were calculated.

3.4. FIS

For the FIS sequences we found no strong strand preference, so we assumed that FIS sites are palindromic in a statistical sense, i.e., each individual FIS binding site need not be palindromic, although the symmetry shows up when averaging over many sites. For these sequences we found a model of length 19 to work well and used it to get the FIS logo presented in *figure 8*, with the consensus site of Finkel and Johnson [20] shown below. Only the flanking G and C in both the consensus and logo plot are well conserved. This is very similar to the logo derived previously [22], although our logo is slightly asymmetric. The sequence conservation is evidently not very strong (i.e., low average information content) suggesting a high degree of non-specificity in the binding of FIS. This is also indicated by *figure 9*, which is a plot analogous to *figure 4*. *Figure 9* shows two things: firstly, the sensitivity level at which false positives start to dominate is about 33% (which means that at best we could predict only about a one-third of the FIS binding sites with confidence); and secondly, the number of positives picked up in a random sequence of genome length is almost the same as for the real genome for all sensitivity levels.

Although we find roughly the same number of FIS sites in the random sequence as in the genome, we do find significant differences in the separations of the found sites, when using the cut-off corresponding to 33% sensitivity. *Figure 10* shows that the genome sites are more clustered than those in the random sequences. Also, as with IHF, we see a higher density of FIS sites immediately upstream of transcription start sites in *figure 11*, which was also expected. This suggests that although the FIS binding is

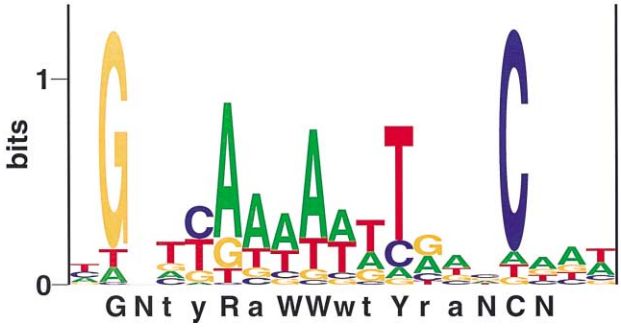

Figure 8. Logo plots for IHF and FIS binding sites in *E. coli*. Shown underneath is the consensus site of Finkel and Johnson [20]. Abbreviations for the nucleotides are: N = A, T, G, or C; R = A or G; Y = C or T; W = A or T. Lower case letters represent less well conserved bases. Note that in the consensus site (as well as in the logo plot above), only the C and G are highly conserved.

fairly non-specific, the model is able to pick out the best binding sites in the *E. coli* genome. The model predicted 944 FIS sites on the direct strand and 938 on the reverse with 239 of these being within 19 bp of each other leaving 1643 separate sites in *E. coli*. If we were to estimate a number, we would guess that there are roughly 6000 strong FIS binding sites in the genome, based on the extrapolation of the linear regime in *figure 9*.

3.5. Global organisation of the E. coli genome

As already mentioned, binding of the FIS protein can be localised to a few certain regions or 'clumps' in the chromosome [2]. *Figure 12* shows that when the FIS binding sites are smoothed with a large window, there are two strong bands flanking the terminus, with a few other less intense bands in the rest of the genome. In contrast, the IHF sites are clumped in many large regions. It is interesting to note that these regions appear to be in between the regions of highly expressed genes. It is possible that the regions of highly expressed genes might reflect topological domains, and that the less-expressed regions (containing a large fraction of IHF protein binding sites) might be domains which are less supercoiled or at any rate require activation of transcription.

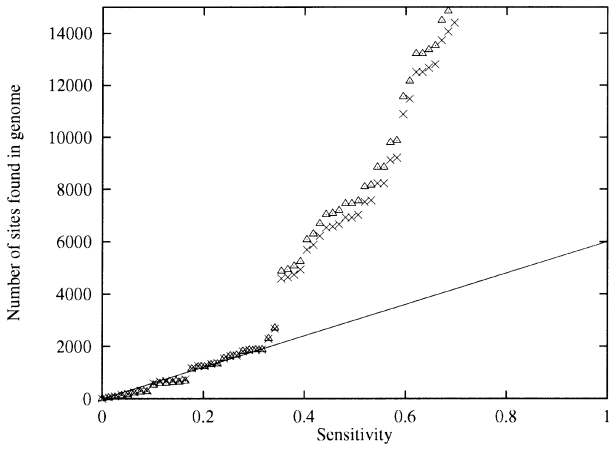

Figure 9. Estimation of the number of FIS sites the complete *E. coli* genome. Same symbols as described in the legend to *figure 4*.

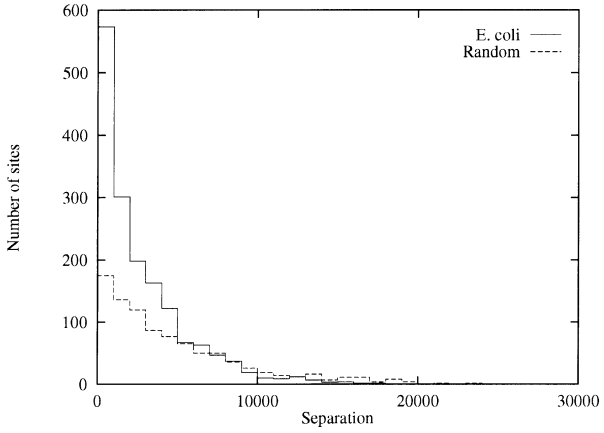

Figure 10. The distribution of separations (measures in base-pairs) between FIS sites in the *E. coli* genome (full line) and, for comparison, in a random sequence of genome length (dashed line).

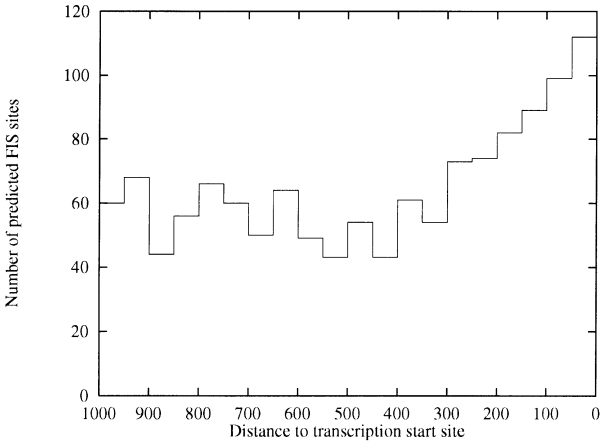

Figure 11. The spatial distribution of FIS binding sites 1000 bases upstream of all transcription starts in *E. coli* genome.

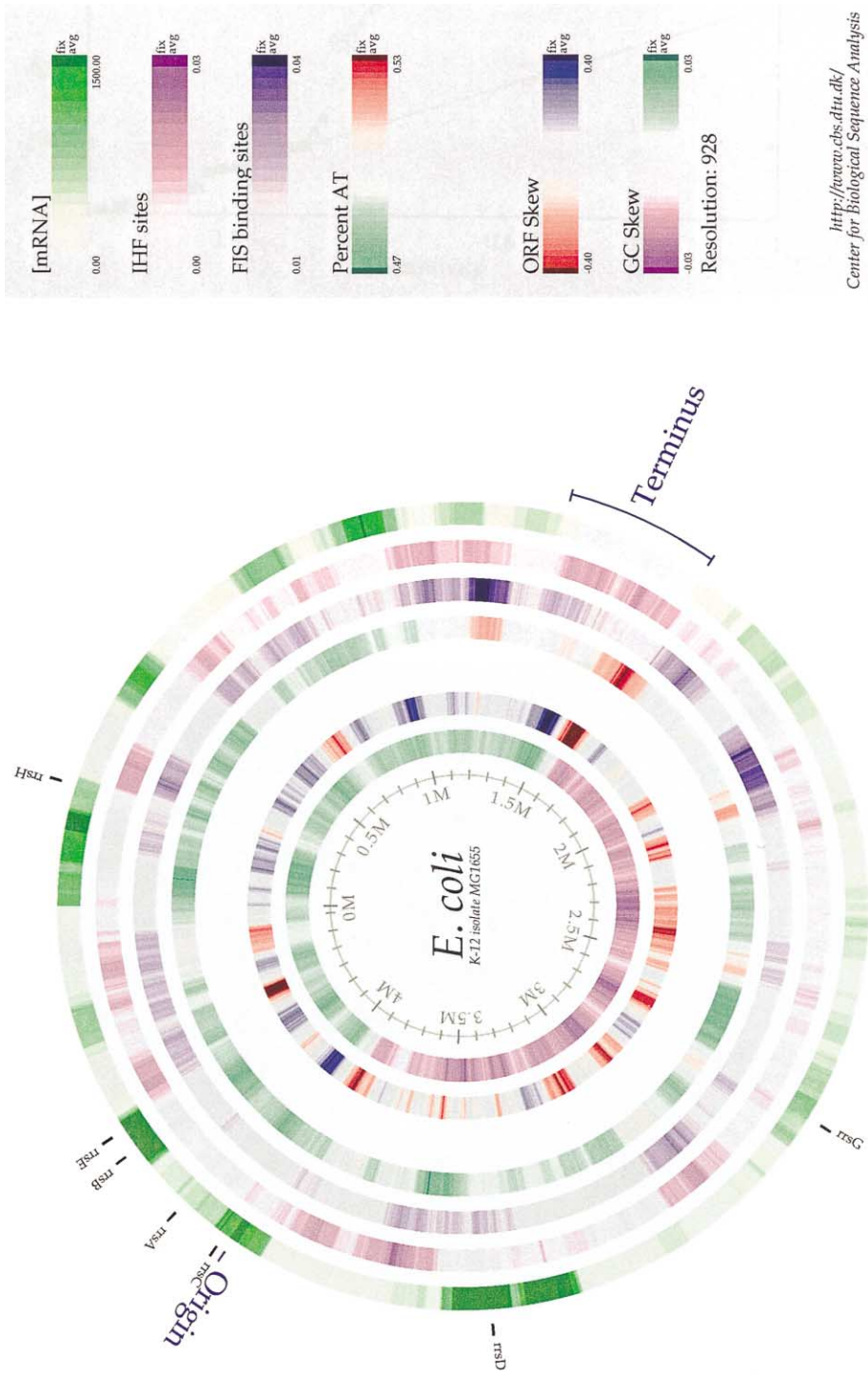

Figure 12. A smoothed version of the GenomeAtlas plot of IHF and FIS binding sites throughout the complete *E. coli* genome. The outer circle (green) is the concentration of mRNA. The next two circles represent the predicted binding sites for IHF and FIS proteins, and the next circle (turquoise/red) is the AT content. The two inner circles represent the skew of the open reading frames (in terms of their direction) and the skew of G's towards one strand or the other.

In addition to the chromatin binding sites, the percent AT of the genome is also plotted in *figure 12*. There is a general trend for the region around the replication terminus to be more AT rich. In addition, there is also a clear skew of direction of open reading frames, with genes (on a global average) tending to orient in the same direction as the leading strand of replication. Finally, the innermost circle shows the GC skew, where the Gs on the leading strand of replication is favoured [25].

In summary, there appear to be several different levels of domains and organisation to the *E. coli* chromosome. These preliminary results hint at a larger global organisation of genes within the *E. coli* genome. We are currently undertaking experiments to explore this role in more detail.

Acknowledgments

This work was supported by a grant from the Danish National Research Foundation. The authors would like to thank Fiona Brew and Garry Miyada of Affymetrix Inc. for kindly providing the *E. coli* Affymetrix chip data prior to publication.

References

[1] Appleman J., Ross W., Salomon J., Gourse R., Activation of *Escherichia coli* rRNA transcription by FIS during a growth cycle, J. Bacteriol. 180 (1998) 1525–1532.

[2] Azam T., Hiraga S., Ishihama A., Two types of localization of the DNA-binding proteins within the *Escherichia coli* nucleoid, Genes Cells 5 (2000) 613–626.

[3] Bachellier S., Clement J., Hofnung M., Short palindromic repetitive DNA elements in enterobacteria: a survey, Res. Microbiol. 150 (1999) 627–639.

[4] Ball C., Osuna R., Ferguson K., Johnson R., Dramatic changes in fis levels upon nutrient upshift in *Escherichia coli*, J. Bacteriol. 174 (1992) 8043–8056.

[5] Betermier M., Galas D., Chandler M., Interaction of FIS protein with DNA: bending and specificity of binding, Biochimie 74 (1994) 958–967.

[6] Bianchi M., The hmb-box domain, in: Lilley D. (Ed.), DNA-Protein: Structural Interactions, IRL Press at Oxford University Press, 1995, pp. 177–200.

[7] Blattner F., et al., The complete genome sequence of *Escherichia coli* K-12, Science 277 (1997) 1453–1474.

[8] Boccard F., Prentki P., Specific interaction of IHF with RIBs, a class of bacterial repetitive DNA elements located at the 3' end of transcription units, EMBO J. 12 (1993) 5019–5027.

[9] Bracco L., Kotlarz D., Kolb A., Diekmann S., Buc H., Synthetic curved DNA sequences can act as transcriptional activators in *Escherichia coli*, EMBO J. 8 (1989) 4289–4296.

[10] Chan S., Breslauer K., Hogan R.A.M., Thermodynamics and premelting conformational changes of phased (da)5 tracts, Biochemistry 32 (1993) 11776–11784.

[11] Choe L., Chen W., Lee K., Proteome analysis of factor for inversion stimulation (Fis) overproduction in *Escherichia coli*, Electrophoresis 20 (1999) 798–805.

[12] Dame R., Wyman C., Goosen N., H-NS mediated compaction of DNA visualised by atomic force microscopy, Nucleic Acids Res. 15 (2000) 3504–3510.

[13] de Vargas L.M., Kim S., Landy A., DNA looping generated by DNA bending protein IHF and the two domains of lambda integrase, Science 244 (1989) 1457–1461.

[14] Dlakic M., Harrington R., The effects of sequence context on DNA curvature, Proc. Natl. Acad. Sci. USA 93 (1996) 3847–3852.

[15] Dorman C., Hinton J., Free A., Domain organization and oligomerization among h-ns-like nucleoid-associated proteins in bacteria, Trends Microbiol. 7 (1999) 124–128.

[16] Dorman C.J., Genetics of Bacterial Virulence, Blackwell Scientific Publications, Oxford, 1994.

[17] Durbin R.M., Eddy S.R., Krogh A., Mitchison G., Biological Sequence Analysis, Cambridge University Press, 1998.

[18] Dworkin J., Ninfa A., Model P., A protein-induced DNA bend increases the specificity of a prokaryotic enhancer-binding protein, Genes Dev. 12 (1998) 894–900.

[19] Engelhorn M., Boccard F., Murtin C., Prentki P., Geiselmann J., In vivo interaction of the *Escherichia coli* integration host factor with its specific binding sites, Nucleic Acids Res. 23 (1995) 2959–2965.

[20] Finkel S., Johnson R., The Fis protein: it's not just for DNA inversion anymore, Mol. Microbiol. 6 (1992) 3257–3265.

[21] Gerard F., Dri A., Moreau P., Role of *Escherichia coli* rpos, lexa and h-ns global regulators in metabolism and survival under aerobic, phosphate-starvation conditions, Microbiology 145 (1999) 1547–1562.

[22] Hengen P., Bartram S., Stewart L., Schneider T., Information analysis of FIS binding sites, Nucleic Acids Res. 25 (1997) 4994–5002.

[23] Hengge-Aronis R., Interplay of global regulators and cell physiology in the general stress response of *Escherichia coli*, Curr. Opin. Microbiol. 2 (1999) 148–152.

[24] Ishihama A., Modulation of the nucleoid, the transcription apparatus, and the translation machinery in bacteria for stationary phase survival, Genes Cells 4 (1999) 135–143.

[25] Jensen L., Friis C., Ussery D., Three views of the *E. coli* genome, Research Microbiol. 150 (1999) 773–777.

[26] Jordi B., et al., DNA binding is not sufficient for H-NS-mediated repression of *proU* expression, J. Biol. Chem. 272 (1997) 12083–12090.

[27] Kostrewa D., et al., Crystal structure of the factor for inversion stimulation FIS at 2.0 a resolution, J. Mol. Biol. 226 (1992) 209–226.

[28] Krogh A., Brown M., Mian I., Sjolande K., Haussler D., Hidden markov models in computational biology. applications to protein modelling, J. Mol. Biol. 235 (1994) 1501–1531.

[29] Link A., Robison K., Church G., Comparing the predicted and observed properties of proteins encoded in the genome of *Escherichia coli* K-12, Electrophoresis 18 (1997) 1259–1313.

[30] Nunes-Duby S., Smith-Mungo L., Landy A., Single base-pair precision and structural rigidity in a small IHF-induced DNA loop, J. Mol. Biol. 253 (1995) 228–242.

[31] Oppenheim A.B., Rudd K.E., Mendelson I., Teff D., Integration host factor binds to a unique class of complex repetitive extragenic DNA sequences in *Escherichia coli*, Mol. Microbiol. 10 (1993) 113–122.

[32] Owen-Hughes T., et al., The chromatin-associated protein H-NS interacts with curved DNA to influence DNA topology and gene expression, Cell 71 (1992) 255–265.

[33] Pedersen A., Jensen L., Staerfeldt H., Brunak S., Ussery D., A DNA structural atlas of *E. coli*, J. Mol. Biol. 299 (2000) 907–930.

[34] Perez-Martin J., de Lorenzo V., Clues and consequences of DNA bending in transcription, Annu. Rev. Microbiol. 51 (1997) 593–628.

[35] Polaczek P., Kwan K., Campbell J., Unwinding of the *Escherichia coli* origin of replication (oriC) can occur in the absence of initiation proteins but is stabilized by DnaA and histone-like proteins IHF or HU, Plasmid 39 (1998) 77–83.

[36] Rice P., Mizuuchi K., Nash H., Crystal structure of an IHF-DNA complex: a protein-induced DNA U-turn, Cell 87 (1996) 1295–1306.

[37] Rohde J., Luan, Rohde H., Fox J., Minnich S., The *Yersinia enterocolitica* pYV virulence plasmid contains multiple intrinsic DNA bends which melt at 37 °C, J. Bacteriol. 181 (1999) 4198–4204.
[38] Schneider T.D., Stephens R.M., Sequence logos: a new way to display consensus sequences, Nucleic Acids Res. 18 (1990) 6097–6100.
[39] Shindo H., et al., Solution structure of the DNA binding domain of a nucleoid-associated protein, H-NS, from *Escherichia coli*, FEBS Lett. 360 (1995) 125–131.
[40] Shpigelman E., Trifonov E., Bolshoy A., CURVATURE: Software for the analysis of curved DNA, CABIOS 9 (1993) 435–444.
[41] Smyth C., et al., Oligomerization of the chromatin-structuring protein h-ns, Mol. Microbiol. 36 (2000) 962–972.
[42] Spassky A., Rimsky S., Garreau H., Buc H., H1a, an *E. coli* DNA-binding protein which accumulates in stationary phase, strongly compacts DNA in vitro, Nucleic Acids Res. 12 (1984) 5321–5340.
[43] Spurio R., et al., Lethal overproduction of the *Escherichia coli* nucleoid protein H-NS: ultramicroscopic and molecular autopsy, Mol. Gen. Genet. 231 (1992) 201–211.
[44] Tanaka K., Muramatsu S., Yamada H., Mizuno T., Systematic characterization of curved DNA segments randomly cloned from *Escherichia coli* and their functional significance, Mol. Gen. Genet. 226 (1991) 367–376.
[45] Thieffry D., Salgado H., Huerta A., Collado-Vides J., Prediction of transcriptional regulatory sites in the complete genome sequence of *Escherichia coli* K-12, Bioinformatics 14 (1998) 391–400.
[46] Travers A., DNA-protein interactions: IHF-the master bender, Curr. Biol. 7 (1997) 252–254.
[47] Travers A., Muskhelishvili G., DNA-microloops and microdomains: a general mechanism for transcription activation by torsional transmission, J. Mol. Biol. 279 (1998) 1072–1073.
[48] Tupper A., et al., The chromatin-associated protein h-ns alters DNA topology in vitro, EMBO J. 13 (1994) 258–268.
[49] Ussery D., Higgins C., Bolshoy A., Environmental influences on DNA curvature, J. Biomol. Struct. Dyn. 16 (1999) 811–823.
[50] Ussery D., et al., The chromatin-associated protein H-NS, Biochimie 76 (1994) 968–990.
[51] van Ulsen P., Hillebrand M., Zulianello L., van de Putte P., Goosen N., The integration host factor-DNA complex upstream of the early promoter of bacteriophage Mu is functionally symmetric, J Bacteriol. 179 (1997) 3073–3075.
[52] VanBogelen R., Olson E., Wanner B., Neidhardt F., Global analysis of proteins synthesized during phosphorus restriction in *Escherichia coli*, J. Bacteriol. 178 (1996) 4344–4366.
[53] Williams R., Rimsky S., Molecular aspects of the *E. coli* nucleoid protein, H-NS: a central controller of gene regulatory networks, FEMS Microbiol Lett 156 (1997) 175–185.
[54] Workman C., Stormo G., Ann-spec: A method for discovering transcription factor binding sites with improved specificity (1999), Proceedings for the Pacific Symposium on Biocomputing, 2000, in press.
[55] Yamada H., Muramatsu S., Mizuno T., An *Escherichia coli* protein that preferentially binds to sharply curved DNA, J. Biochem. (Tokyo) 108 (1990) 420–425.
[56] Yuan H., et al., The molecular structure of wild-type and a mutant Fis protein: relationship between mutational changes and recombinational enhancer function or DNA binding, Proc. Natl. Acad. Sci. USA 88 (1991) 9558–9562.
[57] Zuber F., Kotlarz D., Rimsky S., Buc H., Modulated expression of promoters containing upstream curved DNA sequences by the *Escherichia coli* nucleoid protein H-NS, Mol. Microbiol. 12 (1994) 231–240.