Superhelically Destabilized Sites in the *E. coli* Genome: Implications for Promoter Prediction in Prokaryotes  
Huiquan Wang and Craig J. Benham*

UC Davis Genome Center, University of California  
One Shields Avenue, Davis, CA 95616

**Abstract:** Stress-induced DNA duplex destabilization (SIDD) analysis exploits the known structural and energetic properties of DNA to predict which sites become susceptible to strand separation under superhelical stress. Experimental results show that this analysis is quantitatively accurate in predicting destabilized sites that occur in transcriptional regulatory regions, matrix/scaffold attachment sites and replication origins. Here we report the results of a SIDD analysis of the complete *E. coli* genome, performed using a new algorithm specific for long genomic DNA sequences. Our results demonstrate that less than 7% of the *E. coli* genome has the propensity to become highly destabilized at physiological superhelical densities. Those SIDD sites with high destabilization potential are statistically significantly associated with divergent and tandem intergenic regions, but not with convergent intergenic or coding regions. More than 80% of the intergenic regions containing experimentally characterized promoters are found to overlap these SIDD sites. Strong SIDD sites are highly enriched in the 5' upstream regions of genes regulating stress responses in *E. coli*, suggesting a possible functional role in their regulation. We discuss the possibility of using SIDD properties for promoter prediction, and potential roles of predicted SIDD sites in transcriptional regulation.

**Introduction**

Genomic DNA encodes not only functional RNAs, but also instructions for regulating gene expression. Deciphering such regulatory information has become an important goal, now that complete genome sequences are known. Compared with gene prediction, progress has been slow in developing computational methods to find regulatory regions. Most promoter prediction algorithms are based on statistical features of the DNA string, and have relatively weak predictive power ([We03], [HC03]). Here we consider a new approach, using a structural property of the DNA that we show is closely associated with prokaryotic promoters.

The topological configuration of genomic DNA within bacteria is known to dynamically respond to physiological and environmental events. For example, the state of supercoiling of the *E. coli* genome varies during the transition from growth phase to stationary phase, with osmotic stress, anaerobic stress, or temperature shocks ([Do96]). The global patterns of gene expression also change under these conditions, as shown by microarray experiments ([Ch03], [Sa03]). Because transcription produces waves of positive and negative supercoils, expression of one gene may affect the expression of its neighbors ([LW87]). Changes of chromosomal superhelicity, and hence of the stresses imposed on the duplex, have been suggested as a possible mechanism for global regulation of gene expression *in vivo* ([WL96], [HB02]).

Separation of the DNA strands in the –10 region of the promoter is a necessary step in

*Corresponding author, Email: eibenham@ucdavis.edu; Tel: 530-754-9647.
transcription initiation. Although this event is mediated by interactions with the σ−factor of the promoter, changes in the stability of the DNA duplex can have profound effects on this process. At equilibrium, strand opening frequencies increase exponentially as the free energy required decreases. Thus, a change of only 5 kcal/mol in the intrinsic stability of the DNA duplex can drive the equilibrium of an opening reaction more than 4,000-fold towards the open state. If this opening is the rate-limiting step in a regulatory process, then even fractional destabilization of the DNA duplex can greatly increase its frequency of occurrence.

Stress induced duplex destabilization (SIDD) analysis predicts the extent to which each base pair in a DNA sequence becomes susceptible to strand opening when the DNA is stressed by superhelicity ([Be93]). This susceptibility is not simply a function of local thermodynamic stability, because opening of any base pair alters the partitioning of the superhelicity throughout the region, and hence affects the opening probabilities of every other base pair. In this way stresses couple together the transition behaviors of every base pair that experiences them.

Predicted SIDD sites have been shown to be preferentially located at transcriptional regulatory loci, including promoters and terminators ([Be96]). In eukaryotes, matrix attached regions and replication origins also are highly destabilized ([Be97] and unpublished data). Numerous experiments have implicated SIDD in mechanisms regulating various specific biological processes that either require, or are facilitated by, DNA duplex destabilization. ([LCB00], [He00], [Mi02]) Thus, a systematic analysis of the SIDD profiles of complete genomes will further enhance our understanding of regulatory events within cells.

Here we report the results of the SIDD analysis of the complete E. coli K12 genome (GenBank file NC_000913, version M54), using a new algorithm specific for long genomic DNA sequences. We analyze the relationships between the locations and extents of destabilization of SIDD sites, and the positions of annotated genes in this genome. We show that predicted SIDD sites are statistically enriched in those intergenic regions which are known or inferred to contain promoters, but not in intergenic regions that are inferred not to contain promoters. Critical genes for E. coli adaptation to environmental stresses are shown to be particularly strongly destabilized in their 5’ upstream regions. We discuss the potential application of SIDD analysis in promoter prediction in prokaryotes, and the possible roles of DNA destabilization in gene regulation.

Methods

Computational methods have been developed to calculate the opening probability \( p(x) \) and the free energy \( G(x) \) needed to guarantee opening of each base pair in a relatively short, (i.e. <10 kb) superhelical DNA ([Be93],[FB99]). The superhelical constraint is the linking difference \( \alpha \), the amount by which the actual linking number \( Lk \) of the domain differs from its relaxed value \( Lk_0 \). This linking difference is partitioned between the change of twist involved in opening \( n \) base pairs, helical interwinding \( T \) of those open regions, and the residual linking difference \( \alpha_r \), that remains to stress the domain, according to the conservation equation:

\[
\alpha = -(n/A) + T + \alpha_r = \text{constant}.
\]

Here \( A = 10.5 \) base pairs per turn is the helicity of unstressed B-form DNA. The total free energy for a state of strand separation in a DNA molecule is given by:
\[
G = G_\tau + G_\rho + G_c = C \sum_{j=1}^{N} n_j \gamma_j^2 + \frac{K}{2} \left( \frac{\alpha + n}{A} - \sum_{j=1}^{N} \frac{n_j \gamma_j^2}{2\pi} \right)^2 + \sum_{j=1}^{N} \left( (a + b_j) n_j - a n_j n_j^{+1} \right),
\]

where \( n_j = 1 \) if base pair \( j \) is open in the state, and \( n_j = 0 \) otherwise, and \( \gamma_j \) is the torsional deformation of each open base pair. A complete derivation of this equation is given in ([FB99]). All the conformational and energy parameters are given their experimentally measured values, so there are no free parameters in this analysis.

At thermodynamic equilibrium a population of identical superhelical DNA molecules is distributed among its available states, with low energy states exponentially more populated than high energy states. In a SIDD analysis this equilibrium distribution can be calculated either approximately or exactly ([Be92], [FB99]). Once it is known, the equilibrium value of any parameter of interest may be evaluated. In this way we find the equilibrium probability \( p(x) \) of opening the base pair at each position \( x \), and the incremental free energy \( G(x) \) require to guarantee separation of each base pair ([Be93]). The plot of \( G(x) \) versus \( x \) is called the destabilization (or SIDD) profile. (See Figure 1). A destabilized region appears in the profile as a site where \( G(x) \) is reduced.

These methods are quantitatively accurate at predicting locations where the DNA duplex is susceptible to separation under superhelical stress. In all cases where experiments have been performed, they make highly accurate predications of locations and extents of opening as functions of base sequence and imposed superhelicity ([Be92], [He00]). Moreover, many of the sites that were predicted to separate under stress have subsequently been experimentally shown to open, both in vitro and in vivo ([Ar97], [Be97], [SBH98], [Po03]).

Recently, a new SIDD algorithm has been developed to analyze long DNA sequences, including complete chromosomes ([BB04]). The sequence is partitioned into windows and each window is analyzed separately. All the windows are chosen to have the same length \( N \). Successive windows are offset by a distance \( d_0 \) so each internal base pair appears in \( W = N/d_0 \) windows. Individual windows are analyzed using the previously developed methods, until the entire sequence has been traversed ([Be92], [BB04]). The final values of the opening probability \( p(x) \) and destabilization energy \( G(x) \) at each position \( x \) are calculated as weighted averages of the values computed from all windows containing that base pair. In the present analysis we use \( N = 5000 \) bp, \( d_0 = 500 \) bp, and relative weights \( w = \{1, 2, 4, 8, 16, 16, 8, 4, 2, 1\} \) for the 10 successive windows containing a given base pair. The analysis was performed at superhelix density \( \sigma = \alpha / L_{Ko} = -0.055 \), a mid-physiological value. Details of the algorithm and its implementation are found in ([BB04]).

A typical SIDD profile is shown in Figure 1, annotated with the genes and ORFs in this area. One sees that coding regions are relatively stable, with \( G(x) > 8 \), while intergenic regions are more highly destabilized (i.e. have smaller \( G(x) \) values). The strongest destabilization occurs in the intergenic region separating the divergently oriented dnaK gene from the yaa1 ORF. This region contains two documented dnaK promoters. Substantial destabilization also occurs in other intergenic regions, especially those that are \( 5' \) to an ORF.

**Results**

**Highly destabilized sites comprise a small fraction of the genome** We first determine the number of sites and base pairs that are destabilized below each integer value of \( G(x) \). The results are shown in Table 1, which gives cumulative statistics at each SIDD level. For example, there are a total of 306114 bp in this genome where \( G(x) \leq 4.0 \) kcal/mol. These
occur in 2629 runs, which together constitute 6.60% of the genome (which is 4,639,221 bp long). Similarly, there are 1053 SIDD sites with $G(x) \leq 1.0$ kcal/mol. containing 101102 bps, about 2.18% of the genome. (These base pairs are also counted in the SIDD level 4 statistics because they satisfy $G(x) \leq 4.0$; in this sense the information in Table 1 is cumulative.) As the SIDD level increases the numbers and average lengths of SIDD sites increases, as does their average G+C content. This table shows that the significantly destabilized sites (i.e. those with $G(x) \leq 4.0$ kcal/mol.) comprise a small fraction of the *E. coli* genome.

![Figure 1. The SIDD profile of a representative 5 kbp region of the *E. coli* genome. The genes and ORFs in this region are denoted above the graph, with the upper level transcribing directly. The two experimentally characterized promoters are also indicated](image)

Table I. Cumulative statistics of SIDD sites in the *E. coli* genome

<table>
<thead>
<tr>
<th>SIDD level</th>
<th>Total number</th>
<th>Total length</th>
<th>Mean</th>
<th>STD</th>
<th>GC (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>692</td>
<td>51215</td>
<td>74.0</td>
<td>42.2</td>
<td>26.3</td>
</tr>
<tr>
<td>1</td>
<td>1053</td>
<td>101102</td>
<td>96.0</td>
<td>56.7</td>
<td>29.6</td>
</tr>
<tr>
<td>2</td>
<td>1448</td>
<td>153994</td>
<td>106.3</td>
<td>66.3</td>
<td>32.2</td>
</tr>
<tr>
<td>3</td>
<td>1959</td>
<td>220784</td>
<td>112.7</td>
<td>76.7</td>
<td>34.3</td>
</tr>
<tr>
<td>4</td>
<td>2629</td>
<td>306114</td>
<td>116.4</td>
<td>82.5</td>
<td>36.2</td>
</tr>
<tr>
<td>5</td>
<td>3540</td>
<td>412618</td>
<td>116.6</td>
<td>85.2</td>
<td>37.8</td>
</tr>
<tr>
<td>6</td>
<td>4962</td>
<td>556036</td>
<td>112.1</td>
<td>87.4</td>
<td>39.2</td>
</tr>
</tbody>
</table>

Strong SIDD sites are preferentially associated with specific classes of intergenic regions

To correlate SIDD sites with genic and intergenic regions, we partitioned the complete *E. coli* genome into four disjoint sets. First, DIV is the set of all intergenic regions that separate divergently oriented genes. These may be inferred to contain promoters, but not terminators. TAN is the set of intergenic regions separating tandemly oriented genes. In prokaryotes TAN regions may contain promoters, or terminators, or both, or neither. CON is the set of intergenic regions separating convergently oriented genes. These may be inferred to contain terminators, but not promoters. Finally, Genes is the set of annotated ORFs and RNA genes. The statistics of these regions are given in Table II below.

To analyze associations between destabilized sites and transcriptional regions within the *E. coli* chromosome, we define a SIDD site to be a set of consecutive base pairs which all have $G(x) \leq 8.0$ kcal/mol. The extent of destabilization of each SIDD site is determined by the minimum value $G_{min}$ of $G(x)$ found within it. We also partition SIDD sites into disjoint
bins at levels 0 through 6 according to their \( G_{\text{min}} \) values. Thus, bin 0 contains all SIDD sites with \( G_{\text{min}} \leq 0.0 \) kcal/mol., while bin 1 contains sites with \( 0 < G_{\text{min}} \leq 1.0 \) kcal/mol., and so on.

<table>
<thead>
<tr>
<th>Region</th>
<th>Total number</th>
<th>Total length</th>
<th>Mean</th>
<th>STD</th>
<th>GC (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Divergent (DIV)</td>
<td>624</td>
<td>154935</td>
<td>247.9</td>
<td>182.9</td>
<td>40.0</td>
</tr>
<tr>
<td>Tandem (TAN)</td>
<td>2405</td>
<td>294257</td>
<td>122.2</td>
<td>141.6</td>
<td>42.1</td>
</tr>
<tr>
<td>Convergent (CON)</td>
<td>555</td>
<td>72769</td>
<td>130.9</td>
<td>175.6</td>
<td>46.5</td>
</tr>
<tr>
<td>Overlap</td>
<td>811</td>
<td>21945</td>
<td>27.0</td>
<td>100.0</td>
<td>48.5</td>
</tr>
<tr>
<td>Gene</td>
<td>4395</td>
<td>4139205</td>
<td>941.8</td>
<td>635.9</td>
<td>51.9</td>
</tr>
</tbody>
</table>

We examine the occurrence of SIDD sites at each of the three types of intergenic regions described above, and at intergenic regions containing documented promoters. As shown in Figure 2, the percentage of SIDD sites overlapping intergenic regions increases from 45.4% at level 4 to 89% at level 0. This is due to an increasingly dense clustering of SIDD sites at TAN and DIV, but not at CON.

There are 305 sites in the GenBank file for this sequence that are annotated as documented promoters. Of these, 43 occur within coding regions, and 262 are located within a total of 189 intergenic regions. All of these regions are either in DIV or TAN; none are in CON. Similarly, there are 471 experimentally identified promoters in the PromEC database. ([He01]). Of these, 195 promoters were located in 151 distinct TAN regions, while 204 were in 134 DIV regions. No promoter in either data set was found in a CON region. Taking together, 84.7% of these documented \textit{E. coli} promoters are located in 285 divergent or tandem intergenic regions; we call this set of intergenic regions DTP. A large majority of these DTPs are highly destabilized; 156 out of the 189 GenBank-annotated DTPs (82.5%) and 233 out of 285 DTPs in the promEC database (81.5%) overlap with SIDD sites having \( G_{\text{min}} \leq 6.0 \) kcal/mol.

![Figure 2](image1.png) ![Figure 3](image2.png)

Promoter determination in prokaryotes has proven to be extremely difficult. Only 10% of the genes in \textit{E. coli} have experimentally characterized promoters. Since promoters are usually located in the 5' upstream regions of genes, we determined all SIDD sites that contain either start or stop sites of genes. The results are shown in Figure 3. As progressively more destabilized sites are considered, the percentage of sites that overlap start positions increases, but the percentage that overlaps stops does not. Given the high density of genes and the large number of short intergenic regions in \textit{E. coli}, a TAN region having a SIDD site that overlaps...
the stop site of a gene may also contain a promoter for the next gene. These observations suggest a potentially important role for destabilized sites in transcriptional initiation for a subset - not all - of the genes in \textit{E. coli}.

The statistical significance of associations of SIDD sites with DIV, TAN, DTP, CON and Gene. We determine the statistical significance of the associations found above by comparing them to the associations that would occur if the SIDD sites were randomly located along the \textit{E. coli} genome. We use a simple Monte Carlo approach to select a random set of regions, uniformly distributed along the genome, whose numbers and lengths are the same as those of the actual SIDD sites at a given level. The number of these random regions that overlap DIV, CON, TAN and DTP was found, as was the number that were internal to genes. This procedure was repeated 1000 times for each SIDD level. The resulting distributions appear normal for each SIDD level and interval type, so the means and standard deviations were estimated. Then a z-score was calculated for each observed association with actual SIDD sites. Figure 4 shows the distributions for SIDD level 2.

![Figure 4](image)

Figure 4. A result from the Monte Carlo test with the numbers of the SIDD sites at level 2 overlapping the DTP, DIV, CON and internal genes

The z-scores for associations with the various types of intergenic regions and with genes are shown in Figure 5 for SIDD levels 0 through 6. The overlap between SIDD sites of all levels and internal regions of genes are far below what is expected at random, with z-scores increasing from -30.4 at level 0 to -9.1 at level 6. (A 99.9% confidence interval corresponds to $-3.5 < z < +3.5$.) Associations of SIDD sites at all levels with TAN regions are much higher than expected at random, with z-scores increasing from +7.31 at level 6 to +27.53 at level 0. Associations between strong SIDD sites (levels 0 to 4) and both DIV and DTP are also much higher than expected at random. In contrast, the associations with CON regions are not significantly different from random at the 99.9% confidence level.

Genes regulating stress responses in \textit{E. coli} have strong upstream SIDD sites. In bacteria, chromosomal supercoiling varies during environmental transitions such as pH changes, osmotic stress, O$_2$ deprivation and starvation. These global superhelicity changes could provide a rapid and efficient cellular mechanism to adjust gene expression levels during environmental stress ([WL96], [HB02]). We have examined the SIDD profiles of genes whose expression levels are known to change in response to stress.
Table III lists representative genes encoding important global regulators that alter gene expression patterns in response to environmental stresses. The expression levels of most of these genes are sensitive to superhelicity. The destabilization levels of their upstream regions vary in parallel with their functional roles. RpoS responds to many types of stress by activating downstream genes ([He96], [He99]). A level 0 SIDD site overlaps its 5’ upstream region. In contrast, only a level 7 SIDD site overlaps the 5’ upstream region of its close relative, the “house keeping” sigma 70 rpoD gene. Gyrase introduces negative supercoils into genomic DNA, while topoisomerase I relaxes them. ([Dr92]) The gyrA and gyrB genes have SIDD sites at levels 2 and 5, respectively, overlapping their 5’ upstream regions, while topA and topB have SIDD sites at levels 6 and 8. The histone-like proteins HU (hupA and hupB) and H-NS are known to be involved in global control of DNA supercoiling during stresses ([Do96], [WR97], [Ba01], [DG02]). The 5’ upstream regions of these genes are highly susceptible to destabilization, with SIDD levels 0 for hupA and hupB, and 1 for H-NS. HimA and himD, the genes encoding IHF, an abundant DNA binding protein that may not be involved in supercoiling control during stress, both have relatively stable 5’ upstream regions (a level 4 SIDD site at himA and a level 6 SIDD site at himD). Global regulators such as fnr, crp and lrp regulate different subsets of genes during stresses through their interactions with sigma factor S ([LL96]). They also have distinct SIDD properties in their 5’ upstream regions. A SIDD site at level 1 overlaps the 5’ upstream region of crp gene and a SIDD site at level 0 for fnr, while a SIDD site at level 6 overlaps the 5’ upstream region of lrp gene.

**Table III.** Upstream SIDD properties of genes encoding global transcriptional regulators

<table>
<thead>
<tr>
<th>Stress</th>
<th>SIDD</th>
<th>Gene</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Osmotic, nutrition</td>
<td>0</td>
<td>rpoS</td>
<td>Sigma S (sigma 38) factor of RNA polymerase, major sigma factor during stationery phase</td>
</tr>
<tr>
<td>starvation, cold</td>
<td>2</td>
<td>gyrA</td>
<td>DNA gyrase, subunit A,</td>
</tr>
<tr>
<td>Same as above</td>
<td>0</td>
<td>hupA</td>
<td>DNA-binding protein HU-alpha (HU-2), plays a role in DNA replication and in rpo translation</td>
</tr>
<tr>
<td>Same as above</td>
<td>0</td>
<td>hupB</td>
<td></td>
</tr>
<tr>
<td>Same as above</td>
<td>1</td>
<td>H-NS</td>
<td>Transcriptional regulator, DNA-binding protein HLP-II, increases DNA thermal stability</td>
</tr>
<tr>
<td>Same as above?</td>
<td>1</td>
<td>crp</td>
<td>Transcriptional regulator, cAMP receptor protein (cAMP-binding family), interacts with RNAP</td>
</tr>
<tr>
<td>Aerobic/anaerobic</td>
<td>0</td>
<td>fnr</td>
<td>Transcriptional regulator of aerobic, anaerobic respiration, osmotic balance (cAMP-binding family)</td>
</tr>
<tr>
<td>Aerobic/anaerobic</td>
<td>0</td>
<td>narX</td>
<td>Sensory histidine kinase in two-component regulatory system with NarL, regulation of anaerobic respiration and fermentation, senses nitrate/nitrite</td>
</tr>
<tr>
<td>Aerobic/anaerobic</td>
<td>0</td>
<td>narQ</td>
<td>Sensory histidine kinase in two-component regulatory system with NarP (NarL), regulates anaerobic respiratory gene expression, senses nitrate/nitrite</td>
</tr>
<tr>
<td>Osmotic shock</td>
<td>0</td>
<td>ompR</td>
<td>response regulator in two-component regulatory system with EnvZ, regulates ompF and ompC expression (OmpR family)</td>
</tr>
</tbody>
</table>

Transcriptional regulation of the adaptive response genes in aerobic/anaerobic metabolism are coordinated by fnr protein and proteins of the two-component signal transduction systems, i.e. NarL/NarX and NarP/NarQ ([LL96]). It is intriguing to see that
only one gene in each of these two component systems has a highly destabilized 5' upstream region. NarX and narQ, which both encode membrane-bound sensor histidine kinases, each have a level 0 SIDD site overlapping their 5' upstream region, while narL and narP have SIDD levels 8 and 7, respectively. Similarly, the ompR/envZ two component signal transduction system regulates transcription of the porin regulon ompF and ompC ([LL96]). A level 0 SIDD site overlaps the 5' upstream region of the regulator ompR gene but only a level 8 SIDD site overlaps the envZ gene region. Although both narL/narX and ompR/envZ gene pairs may transcribe as operons, with narX and ompR the first gene respectively. The narP/narQ gene pair are far apart in the genome. In all, 24 of the 28 reported two-component signal transduction systems have at least one SIDD site below level 4 overlapping the upstream regions of the gene for either the sensor or the regulator ([Mi97]). Taken together, these data suggest upstream regions of genes which regulate stress responses in E. coli are intrinsically highly destabilized. This may suggest that regulated destabilization at these sites may play some role in their mechanisms of function.

Discussion

Here we report the SIDD analysis of the complete E. coli K-12 genome. Strong SIDD sites are statistically significantly associated with intergenic regions that contain documented promoters, and with divergent and tandem regions which may contain promoters or operators, while SIDD sites internal to genes are statistically highly underrepresented. These results agree with previous analyses of small numbers of individual genes ([Be93], [Be96]). In both yeast and E. coli we find a pattern in which destabilized sites are strongly concentrated in intergenic regions and avoid coding regions. However, in yeast the strongest SIDD sites are in 3' gene flanks, whereas in E. coli they are in 5' flanks. The reason for this difference is not known.

The extreme degree of clustering of SIDD sites at promoters is shown by the fact that the density of strongest SIDD sites is eighty five times greater in promoter-containing intergenic regions than in genes! This suggests that SIDD properties could be used to predict which intergenic regions contain promoters in E. coli, and perhaps also in other prokaryotes, a problem that has proven surprisingly difficult to resolve using string-based methods alone. ([HS96], [VMS99], [Es03]) One such method is based on the hexameric sequence properties of putative -10 and -35 regions. ([HC03]) The best implementation of this method achieves an overall accuracy of 53%. We contrast this with our results, which are based on SIDD properties alone. If a SIDD site with \( G_{\text{min}} \leq 2.0 \text{ kcal/mol} \) overlaps an intergenic region, our results can be used to estimate the probability that this region contains a promoter to be approximately 80%, while if it does not this probability falls below 25%. Suppose one were to predict the presence of a promoter within an intergenic region based exclusively on whether or not it overlaps a destabilized site at this level. Using the known number of intergenic regions that contain such SIDD sites, we estimate this predictor to have a 67% statistical accuracy. Thus, predictions that are based exclusively on this one structural attribute would substantially outperform even the best current sequence-based predictor.

We do not intend to suggest that all predicted SIDD sites overlap actual promoters; instead we propose that many of them are spatially close to promoters, and may play active roles in mechanisms of gene expression. For example, a strong SIDD site is located in the 5' upstream region of the ilvP\(_G\) promoter. It overlaps an IHF binding site, but not the promoter. When IHF binds it closes this SIDD region. The negative superhelical twist normally absorbed by separation at the SIDD site is then transferred to the -10 region of the ilvP\(_G\) promoter ([SBH98]). This binding-induced transmission of destabilization illustrates the
active role that strand opening can play in regulation. We note that this mechanism does not require SIDD at the promoter, but rather at a regulatory protein binding site.

Chromosomal superhelicity can change quickly when a bacterium encounters an environmental stress ([Hi88], [LF00], [Re03]). If this change acts to adjust gene expression levels, one would expect it to regulate the genes that are the global stress regulators. These would be the upstream genes in regulatory systems, most of which are known to be sensitive to both environmental stress and DNA supercoiling. Our results suggest that genes encoding proteins that function under different conditions may have distinctive SIDD properties, an example being the sigma factors rpoS and rpoD (Table III). The proteins in two component signal transduction systems monitor environmental stimulations and mediate changes in gene expression or cell behavior accordingly. Of the 28 pairs that have been identified to date, the gene pairs encoding 24 have at least one SIDD site below level 4 overlapping either the sensory kinase gene or the response regulator gene. The preliminary analysis performed to date suggests that genes encoding either transcription factors or environmental stress response elements have strong SIDD sites in their 5’ flanks. One might evaluate this by determining the upstream SIDD properties of those genes whose expression levels are found by microarray analysis to change. While key response regulatory genes were often associated with strong SIDD sites, we note that microarray experiments detect expression levels after adaptation to the new environment. Thus, they do not isolate the responses of the initial regulatory genes. Nonetheless, our results suggest that strong SIDD sites are not randomly distributed among the genes in E. coli, but rather are associated with certain subsets of genes with distinct functions, in particular regulatory functions.

We are currently analyzing SIDD profiles for other bacterial genomes to validate and extend our findings in E. coli.

References


