Nucleic Acids Research, 2006, Vol. 34, Database issue ***D373–D378***
doi:10.1093/nar/gkj007

SIDDBASE: a database containing the stress-induced DNA duplex destabilization (SIDD) profiles of complete microbial genomes

Huiquan Wang, Miroslava Kaloper and Craig J. Benham*

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

Received July 28, 2005; Revised and Accepted September 12, 2005

ABSTRACT

Prokaryotic genomic DNA is generally negatively supercoiled *in vivo*. Many regulatory processes, including the initiation of transcription, are known to depend on the superhelical state of the DNA substrate. The stresses induced within DNA by negative superhelicity can destabilize the DNA duplex at specific sites. Various experiments have either shown or suggested that stress-induced DNA duplex destabilization (SIDD) is involved in specific regulatory mechanisms governing a variety of biological processes. We have developed methods to evaluate the SIDD properties of DNA sequences, including complete chromosomes. This analysis predicts the locations where the duplex becomes destabilized under superhelical stress. Previous studies have shown that the SIDD-susceptible sites predicted in this way occur at rates much higher than expected at random in transcriptional regulatory regions, and much lower than expected in coding regions. Analysis of the SIDD profiles of 42 bacterial genomes chosen for their diversity confirms this pattern. Predictions of SIDD sites have been used to identify potential genomic regulatory regions, and suggest both possible regulatory mechanisms involving stress-induced destabilization and experimental tests of these mechanisms. Here we describe the SIDDBASE database which enables users to retrieve and visualize the results of SIDD analyses of completely sequenced prokaryotic and archaeal genomes, together with their annotations. SIDDBASE is available at www.gc.ucdavis.edu/benham/siddbase.

INTRODUCTION

In vivo, the biologically active form of DNA in prokaryotes is negatively supercoiled. The amount of superhelical stress imposed on the DNA is determined by the levels of competing DNA topoisomerase enzyme activities, and by local events such as protein binding or DNA transcription (1–3). Transient changes in the level of global DNA supercoiling have been observed with several types of environmental stress, including heat shock, cold shock, pH changes, osmotic shifts, transitions from aerobiosis to anaerobiosis and starvation (1,2). Along with these changes in stress level, the expression patterns of the bacteria involved were observed to be dramatically altered (4–6). Reactions occurring on the DNA template, including transcription and replication, also affect the local level of supercoiling. When RNA polymerase threads through the DNA template, it pushes a wave of positive supercoils ahead, and leaves a trail of negative supercoils behind (3,7). This can affect the expression of nearby genes.

One way in which negative superhelicity can influence regulation is through the destabilizing effect it has on the double helix at susceptible locations within the sequence. Destabilization by even a few kilocalories, far less than would be required to open the duplex, can have a profound effect on the ability of a regulatory molecule to unpair the DNA, as is required for the initiation of transcription or replication. In this manner the modulation of superhelicity can affect both local and global patterns of gene expression. Experiments have shown that stress-induced duplex destabilization plays essential roles in the transcriptional regulation of several genes (8–10).

We have developed computational methods that evaluate the patterns of stress-induced DNA duplex destabilization (SIDD) in DNA sequences (11,12). These analyses predict the locations where the DNA duplex becomes susceptible to separation when under superhelical stress. All conformational and thermodynamic parameters are given their experimentally measured values, so there are no free parameters in these analyses. Despite this, their results are in quantitative agreement with experiments in all cases where experimental information is available.

When the entire *Escherichia coli* genome is analyzed in this way, the sites that are predicted to be easily stress-destabilized

*To whom correspondence should be addressed. Tel: +1 530 754 9647; Fax: +1 530 754 9658; Email: cjbenham@ucdavis.edu

are found not to be distributed at random. Instead, these SIDD sites are highly enriched in those intergenic regions that are known or inferred to contain promoters, and occur infrequently in coding regions (13). Both components of this pattern have very high statistical significances. The frequency of SIDD sites in intergenic regions separating convergent open reading frames (ORFs), which are inferred not to contain promoters, are consistent with random. A similar pattern of SIDD sites avoiding coding regions and being enriched in intergenic regions was noted in yeast, although there the strongest SIDD sites were in the terminal flanks of genes, not in their promoters (14). Our most recent studies also indicate that those genes in *E.coli* whose promoters have strong SIDD sites are clustered in certain functional groups such as transcription regulators, transport and membrane proteins. It is interesting that many known supercoil-responsive genes and environmental stress-responsive genes have highly destabilized sites in their upstream 5′ flanks.

SIDD sites have been shown to be important functional elements in regulating transcriptional initiation, transcriptional termination and other biological activities. In *E.coli*, activation of both the ilvPG and leuV promoters are mediated by similar mechanisms involving a binding-induced translocation of superhelical tension from a SIDD site to the promoter (8,9). This translocated superhelical tension facilitates the formation of the open initiation complex by unwinding the DNA duplex in the promoter region. In humans, the initiation of transcription of the c-myc gene is regulated by the binding of FBP to a highly destabilized SIDD site (15). SIDD sites also have been implicated in transcriptional termination and

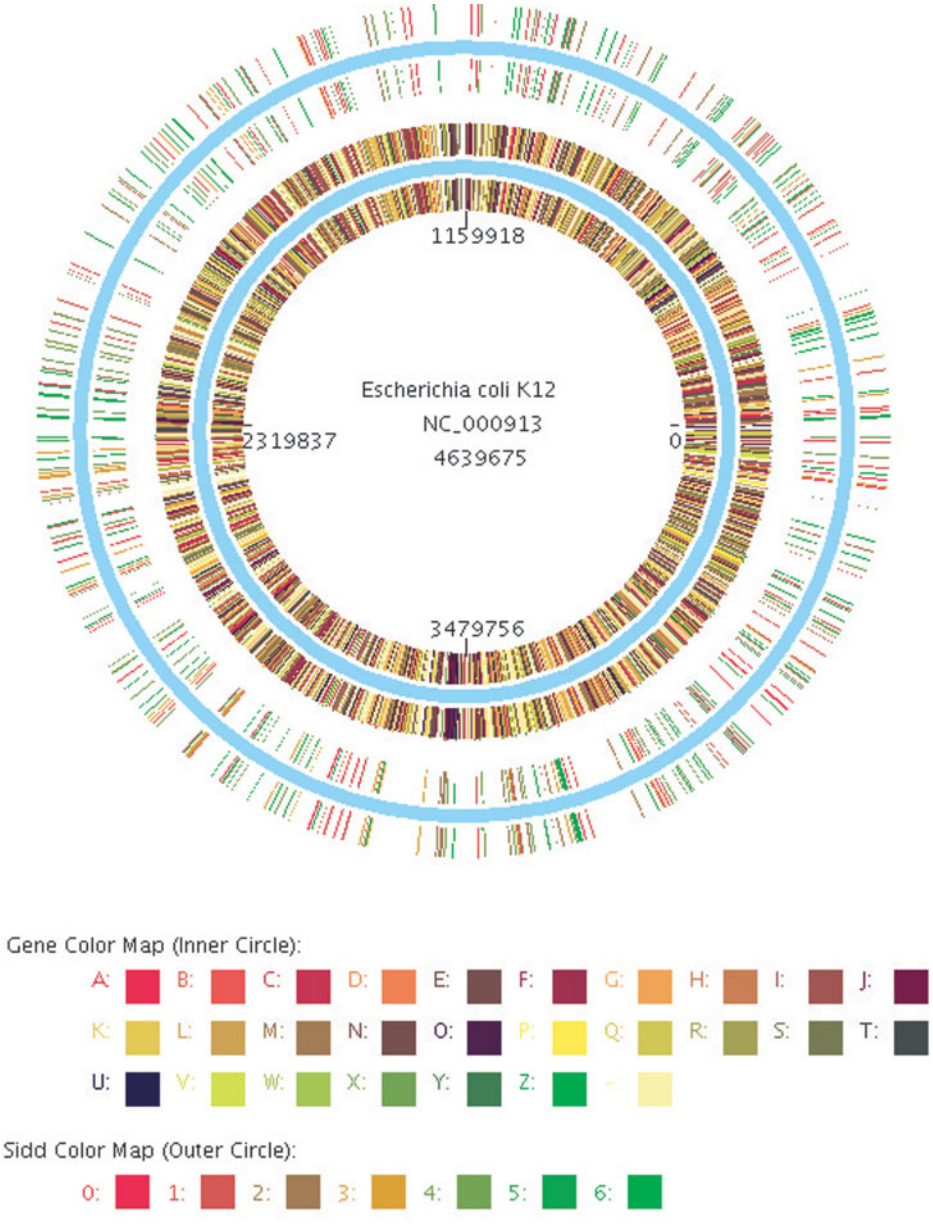

Figure 1. An overview of SIDD sites in the *E.coli* genome. The genes are plotted in the inner circle, color coded according to their COG classifications as shown in the gene color map. The SIDD sites are shown in the outer circle, color coded according to their minimum SIDD energy as shown in the SIDD color map.

chromosomal matrix attachment in yeast (16). These and other results show that SIDD is an essential component of regulatory mechanisms for a variety of biological activities.

It is important to understand that SIDD properties are not simply reflections of the underlying thermal stability of the sequences involved. Stresses couple together the destabilization behaviors of all base pairs that experience them. This leads to much more complex, interactive behaviors than that occur with thermal melting. Specific examples have been presented elsewhere. [See Figures 1 and 2 of Ref (17).]

Here we describe the database we are compiling of the SIDD profiles of microbial genomes. Accessible over the web at www.gc.ucdavis.edu/benham/siddbase, it gives users an overview of the SIDD sites in their selected genome, and their positions relative to the annotated genes. This information will facilitate the identification of regulatory elements, such as promoter-containing regions, in the genomic sequence. In addition, the original SIDD profiles (raw data and graphs) can be visualized and made available for downloading.

MATERIALS AND METHODS

The Refseq sequences of the analyzed microbial genomes were downloaded from the NIH microbial website http://www.ncbi.nlm.nih.gov/genomes/MICROBES/Complete.html. In all cases the date of last revision was not earlier than June 24, 2004. The gene annotation information for each genome was extracted from its GenBank file. The protein gene products are classified into functional categories according to the information in the Clusters of Orthologous Groups (COG) database (18).

Our research group has developed three algorithmic strategies to evaluate the equilibrium distribution of states of destabilization of a short DNA sequence in response to

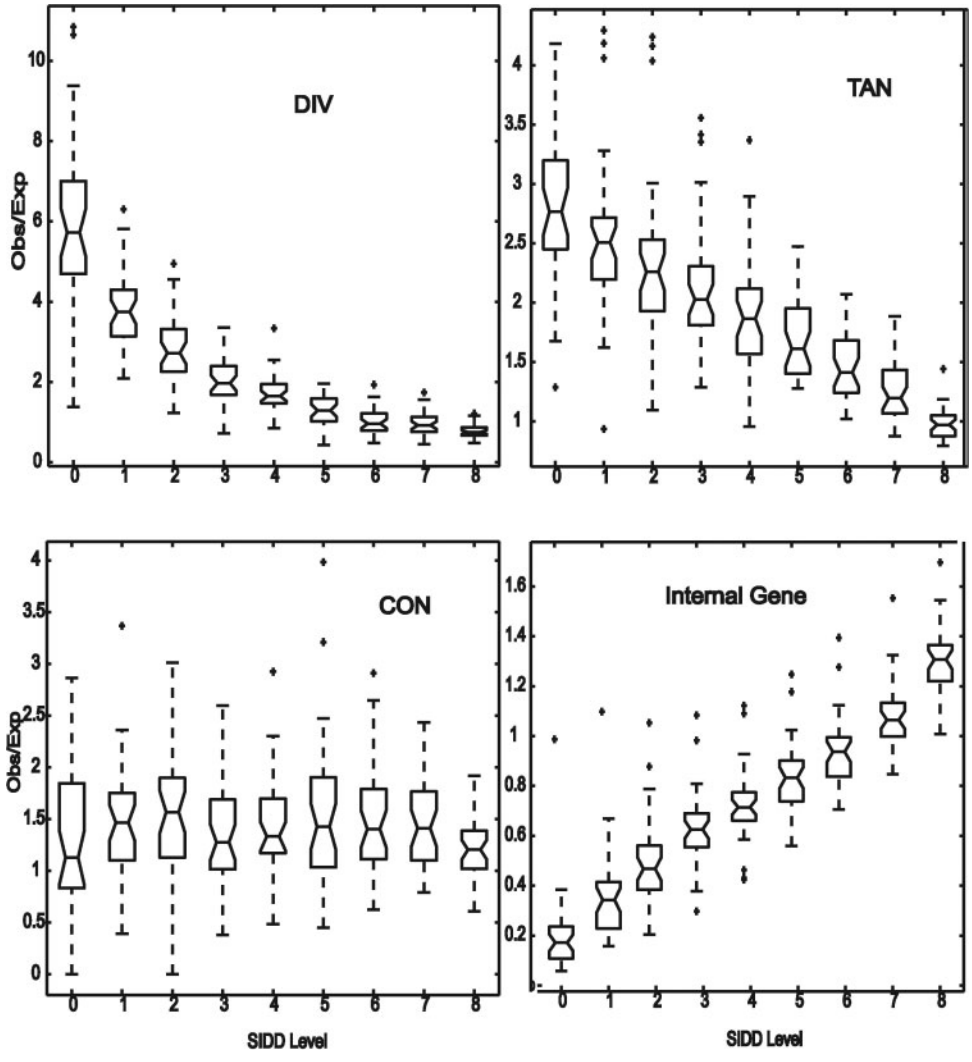

Figure 2. The ratio of the observed number of SIDD sites to the number expected if they were located at random. This ratio has been calculated for SIDD sites that overlap intergenic regions whose flanking ORFs are in any of three orientations (DIV, divergent; CON, convergent; and TAN, tandem), and also for those that occur within coding regions. These results were calculated from the SIDD profiles of 42 bacterial genomes that were chosen to represent the phylogenetic diversity of sequenced genomes and the range of AT/GC ratios. The *X*-axis is the SIDD level; the *Y*-axis is the ratio of predicted SIDD sites found in the regions to the expected number of such sites if they were located at random. This shows that the pattern reported previously in *E.coli* K12 occurs throughout the sequenced prokaryotes.

negative superhelicity (12,19,20). These methods can calculate the probability of opening of each base pair in the sequence. One algorithm also calculates the incremental free energy $G(x)$ needed to guarantee opening of the base pair at position x. This can be done for each base pair in the sequence. Strongly destabilized sites require little or no extra free energy to open, so their values of $G(x)$ are near zero. Sites that remain virtually as stable as they would be in relaxed conditions (which is the majority of the genome) have $G(x)$ near 10 kcal/mol. Partially destabilized sites have intermediate values.

These methods have been extended recently to enable the analysis of long DNA sequences, and successfully applied to the complete genome of *E.coli* (13,17). (For detailed information on the algorithms and the methods for analyzing their results, please refer to the cited publications.) The SIDD analysis of complete microbial genomes has been semi-automated on a 38 node Apple cluster. All SIDD profiles were calculated at superhelical density $\sigma = \Delta Lk/Lk_0 = -0.06$, a moderate physiological value. The results from the calculations were manually reviewed for integrity, then their global characteristics were analyzed by a set of Perl scripts and C++ programs. The results were directly channeled to a PostGres database for storage, visualization and further analysis.

RESULTS

As of the present (September 8, 2005) there are 134 analyzed microbial genomes in this database, 118 from bacteria and 16 from archaea. We update the database whenever the SIDD analysis of another fully sequenced genome is completed, so the number of analyzed genomes will rise in the future. For each analyzed genome, the pattern of global destabilization can be visualized as shown in Figure 1. Each SIDD site in the graph is a set of contiguous base pairs for all of which $G(x) < 8$ kcal/mol. These SIDD sites were binned into disjoint sets according to the minimum value G_m that $G(x)$ attains within them. The lowest bin is determined by $G_m \leqslant 0$, and the other bins contain the SIDD sites satisfying $i-1 < G_m \leqslant i$ for $i = 1, \ldots, 6$. A color map scheme is used to represent these binned SIDD sites.

The SIDD sites in all the fully analyzed bacterial genomes have a similar pattern of distribution to that reported previously for the *E.coli* genome (13). Strong destabilization preferentially occurs in the intergenic regions separating divergently (DIV) or tandemly (TAN) transcribed ORFs, while avoiding coding regions. Destabilization in intergenic regions separating convergently transcribing ORFs (CON), which may be inferred not to contain promoters, is consistent with random. This trend is clearly demonstrated in Figure 2, which summarizes data from the analysis of SIDD locations in 42 bacterial genomes. For each genome in the SIDDBASE database we provide a table summarizing the number of SIDD sites at each level of destabilization, and the number of these that occur in the three types of intergenic regions DIV, TAN and CON. Systematic analysis is underway to compare the SIDD properties of different strains of the same species, between different phylogenetic groups of the same kingdom and between different kingdoms.

One also can display detailed SIDD information for any specific region of interest, as shown in Figure 3. These requests can be made by clicking on a region of the circular map, or by specifying the site either by identifying an annotated gene it contains or by its chromosomal location. The figure displayed is a plot of the oriented genes in a 10 kb window centered on the requested position, together with the locations of the SIDD sites in the region. The genes are labeled, their orientations are

Figure 3. A segment of genomic DNA with annotated SIDD sites and genes. This view is obtained from the window of Figure 1 by clicking on a gene, or by entering a gene name or chromosomal location into the appropriate field. The genes are annotated, and the SIDD sites are displayed as colored bars below the line.

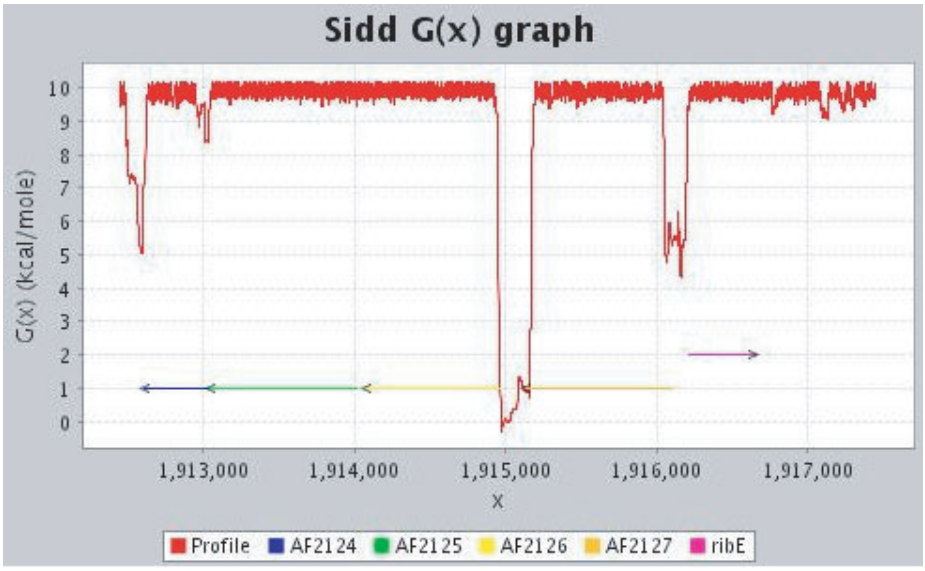

Figure 4. Clicking on a SIDD color bar gives an annotated SIDD profile centered on that region, as shown. From this window one can download the plotted SIDD data for further analysis offline.

shown by arrowheads, and they are color coded according to their COG classification. (RNA genes are put in a special category 'X'.) The end of an arrowhead corresponds to the stop codon position of a gene. The SIDD sites are shown as colored bars, coded according to their G_m values, and displayed below the line showing the annotated genes.

The graph of the SIDD profile of a 5 kb long region centered on a SIDD site can be viewed by clicking on the color bar corresponding to that site. An example is shown in Figure 4. The annotated genes in this area are identified and shown as colored arrows. The SIDD profile data used to generate each such graph can be downloaded directly.

While the SIDD calculation for an entire bacterial genome is time consuming, calculations for short DNA sequences (viz 5 kb) can be executed efficiently. We have provided a website where users can calculate SIDD profiles of short sequences of interest to them. This site may be accessed at http://www.genomecenter.ucdavis.edu/benham/sidd (21). There one can set some of the calculation parameters, including the assumed superhelix density. It should be noted that the results calculated from the web server may not necessarily be identical as the ones displayed on this database, even when the sequences and the parameters are the same. The SIDD profile of a DNA segment in this database was calculated in its native global genomic context, while the results from the web server calculation were not.

FUTURE DEVELOPMENTS

The current database only contains SIDD profiles of complete genomes from prokaryotes and archaea. In the future we also will deposit the results of SIDD analyses for eukaryotic genomes. Initially this will be yeast, at least one complete chromosome from each fully sequenced eukaryote, and the ENCODE regions of the human genome. We intend eventually to include the SIDD profiles of the complete genomes of all fully sequenced model organisms. We will also provide SIDD profiles of specific prokaryotic genomes at several superhelical densities. We will add further functionalities to the database as these are developed, including SIDD-based (or SIDD-assisted) promoter predictions. We will continue to analyze more microbial genomes as their sequences are completed, and we will periodically update our analyses in response to significant changes in NCBI Refseq sequences.

ACKNOWLEDGEMENTS

This work was supported in part by grants to CJB from the National Science Foundation (DBI-0416764) and the National Institutes of Health (RO1-GM68903). Funding to pay the Open Access publication charges for this article was provided by the National Science Foundation.

Conflict of interest statement. None declared.

REFERENCES

1. Drlica,K. (1992) Control of bacterial DNA supercoiling. *Mol. Microbiol.*, **6**, 425–433.
2. Dorman,C.J. (1996) Flexible response: DNA supercoiling, transcription and bacterial adaptation to environmental stress. *Trends Microbiol.*, **4**, 214–216.
3. Wang,J.C. and Lynch,A.S. (1996) Effects of DNA Supercoiling on Gene Expression. In Lin,E.C.C. and Lynch,A.S. (eds), *Regulation of Gene Expression in Escherichia coli*. R.D. Landes & Co., Austin TX, pp. 127–147.
4. Chang,D.E., Smalley,D.J. and Conway,T. (2002) Gene expression profiling of *Escherichia coli* growth transitions: an expanded stringent response model. *Mol. Microbiol.*, **45**, 289–306.
5. Cheung,K.J., Badarinarayana,V., Selinger,D.W., Janse,D. and Church,G.M. (2003) A microarray-based antibiotic screen identifies a regulatory role for supercoiling in the osmotic stress response of *Escherichia coli*. *Genome Res.*, **13**, 206–215.
6. Salmon,K., Hung,S.P., Mekjian,K., Baldi,P., Hatfield,G.W. and Gunsalus,R.P. (2003) Global gene expression profiling in *Escherichia*

coli K12. The effects of oxygen availability and FNR. *J. Biol. Chem.*, **278**, 29837–29855.
7. Liu,L.F. and Wang,J.C. (1987) Supercoiling of the DNA template during transcription. *Proc. Natl Acad. Sci. USA*, **84**, 7024–7027.
8. Sheridan,S.D., Benham,C.J. and Hatfield,G.W. (1998) Activation of gene expression by a novel DNA structural transmission mechanism that requires supercoiling-induced DNA duplex destabilization in an upstream activating sequence. *J. Biol. Chem.*, **273**, 21298–21308.
9. Opel,M.L., Aeling,K.A., Holmes,W.M., Johnson,R.C., Benham,C.J. and Hatfield,G.W. (2004) Activation of transcription initiation from a stable RNA promoter by a Fis protein-mediated DNA structural transmission mechanism. *Mol. Microbiol.*, **53**, 665–674.
10. Kouzine,F., Liu,J., Sanford,S., Chung,H.J. and Levens,D. (2004) The dynamic response of upstream DNA to transcription-generated torsional stress. *Nature Struct. Mol. Biol.*, **11**, 1092–1100.
11. Benham,C.J. (1979) Torsional stress and local denaturation in supercoiled DNA. *Proc. Natl Acad. Sci. USA*, **76**, 3870–3874.
12. Benham,C.J. (1993) Sites of predicted stress-induced DNA duplex destabilization occur preferentially at regulatory loci. *Proc. Natl Acad. Sci. USA*, **90**, 2999–3003.
13. Wang,H., Noordewier,M. and Benham,C.J. (2004) Stress-induced DNA duplex destabilization (SIDD) in the *E.coli* genome: SIDD sites are closely associated with promoters. *Genome Res.*, **14**, 1575–1584.
14. Benham,C.J. (1996) Duplex destabilization in superhelical DNA is predicted to occur at specific transcriptional regulatory regions. *J. Mol. Biol.*, **255**, 425–434.
15. He,L., Liu,J., Collins,I., Sanford,S., O'Connell,B., Benham,C.J. and Levens,D. (2000) Loss of FBP function arrests cellular proliferation and extinguishes c-myc expression. *EMBO J.*, **19**, 1034–1044.
16. Benham,C., Kohwi-Shigematsu,T. and Bode,J. (1997) Stress-induced duplex DNA destabilization in scaffold/matrix attachment regions. *J. Mol. Biol.*, **274**, 181–196.
17. Benham,C.J. and Bi,C. (2004) The analysis of stress-induced duplex destabilization in long genomic DNA sequences. *J. Comput. Biol.*, **11**, 519–543.
18. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
19. Sun,H.Z., Mezei,M., Fye,R. and Benham,C.J. (1995) Monte Carlo analysis of conformational transitions in superhelical DNA. *J. Chem. Phys.*, **103**, 8653–8665.
20. Fye,R.M. and Benham,C.J. (1999) Exact method for numerically analysis a model of local denaturation in superhelically stressed DNA. *Phys. Rev.*, **E59**, 3408–3426.
21. Bi,C. and Benham,C.J. (2004) WebSIDD: server for prediction stress-induced duplex destabilized (SIDD) sites I superhelical DNA. *Bioinformatics*, **20**, 1477–1479.