Genome-wide prediction of G4 DNA as regulatory motifs: Role in *Escherichia coli* global regulation

Pooja Rawal,[1] Veera Bhadra Rao Kummarasetti,[1] Jinoy Ravindran,[1] Nirmal Kumar,[1] Kangkan Halder,[2] Rakesh Sharma,[1,4] Mitali Mukerji,[1,3] Swapan Kumar Das,[3] and Shantanu Chowdhury[1,2,5]

[1]G.N. Ramachandran Knowledge Centre for Genome Informatics, [2]Proteomics and Structural Biology Unit, [3]Functional Genomics Unit, [4]Environmental Biotechnology Unit, Institute of Genomics and Integrative Biology, CSIR, Delhi 110 007, India

The role of nonlinear DNA in replication, recombination, and transcription has become evident in recent years. Although several studies have predicted and characterized regulatory elements at the sequence level, very few have investigated DNA structure as regulatory motifs. Here, using G-quadruplex or G4 DNA motifs as a model, we have researched the role of DNA structure in transcription on a genome-wide scale. Analyses of >61,000 open reading frames (ORFs) across 18 prokaryotes show enrichment of G4 motifs in regulatory regions and indicate its predominance within promoters of genes pertaining to transcription, secondary metabolite biosynthesis, and signal transduction. Based on this, we predict that G4 DNA may present regulatory signals. This is supported by conserved G4 motifs in promoters of orthologous genes across phylogenetically distant organisms. We hypothesized a regulatory role of G4 DNA during supercoiling stress, when duplex destabilization may result in G4 formation. This is in line with our observations from target site analysis for 55 DNA-binding proteins in *Escherichia coli*, which reveals significant ($P < 0.001$) association of G4 motifs with target sites of global regulators FIS and Lrp and the sigma factor RpoD (σ^{70}). These factors together control >1000 genes in the early growth phase and are believed to be induced by supercoiled DNA. We also predict G4 motif-induced supercoiling sensitivity for >30 operons in *E. coli*, and our findings implicate G4 DNA in DNA-topology-mediated global gene regulation in *E. coli*.

[Supplemental material is available online at www.genome.org and http://www.igib.res.in/prokaryote/PG4.htm.]

DNA adopts several secondary structure motifs, although the Watson-Crick duplex is its predominant natural state in genomes. The role of non-B DNA motifs in recombination, replication, and particularly, regulation of gene expression has been implicated and generally appreciated in recent years, although still relatively less understood (Sinden 1994; Perez-Martin and de Lorenzo 1997; Pedersen et al. 2000; Bacolla and Wells 2004). It is now evident that DNA sequence also encodes for spatial structures, much like protein sequence, apart from protein-coding and *cis*-acting regulatory elements. Cells use these structural motifs in such a way that DNA sequence information per se has a minimal role other than facilitating formation of the structural motifs. Several reports have implicated the role of non-B DNA structures in the context of gene regulation, both in prokaryotes (for review, see Hatfield and Benham 2002) and eukaryotes (for review, see Rich and Zhang 2003; Bacolla and Wells 2004). Many studies have predicted and determined regulatory elements at the sequence level (Wasserman et al. 2000; Beer and Tavazoie 2004; Xie et al. 2005); however, very few have investigated DNA structure in this context (Florquin et al. 2005). We have focused on searching and investigating the role of a particular type of non-B DNA motif—the G-quadruplex or G4 DNA as a structural regulatory signal.

Guanine-rich sequences attain unique four-stranded conformations known as G4 DNA (Gellert et al. 1962; Sen and Gilbert 1988; Balagurumoorthy and Brahmachari 1994). G4 DNA stabilized by charge coordination with monovalent cations (especially K^+) within a planar array of four-hydrogen-bonded guanines (G-quartets or tetrads) may result from intramolecular or intermolecular association of four DNA strands in parallel or antiparallel orientation (Fig. 1; for review, see Gilbert and Feigon 1999). Chromosomal regions containing guanine-rich sequence capable of forming G4 DNA include immunoglobin heavy-chain switch regions (Dunnick et al. 1993), G-rich minisatellites (Weitzmann et al. 1997), rDNA (Hanakahi et al. 1999), and telomeres (Parkinson et al. 2002). G4 DNA has been implicated in regulation of the human oncogene *c-myc* (Siddiqui-Jain et al. 2002; Seenisamy et al. 2004) and as an "at-risk motif" involved in genome rearrangements in the nematode *Caenorhabditis elegans* (Cheung et al. 2002).

In vivo structure formation by DNA may have deleterious consequences as established by human neurodegenerative diseases caused by triplet repeat expansions (McMurray 1999; Sinden 1999; Cummings and Zoghbi 2000). Furthermore, non-B DNA structures are targeted by the cellular mismatch repair factors, wherein any lacking factors cause repeat instability in *Saccharomyces cerevisiae* (Strand et al. 1993) and tumors in humans (Kolodner 1995; Modrich and Lahue 1996). DNA secondary structures, particularly G4 DNA, also play a central role in telomere extension and are the focus of targeted anticancer drug development (Zahler et al. 1991; Neidle and Read 2000; Incles et al. 2004). It is known that the *Escherichia coli* RecQ can unwind G4 DNA and that the family of RecQ helicases is conserved and is essential for genomic stability in organisms from *E. coli* to humans (Shen and Loeb 2000; Wu and Maizels 2001; Bachrati and Hickson 2003). However, no systematic investigation of G4 DNA in prokaryotes exists, except one recent study showing in vivo existence of G4 DNA in *E. coli* (Duquette et al. 2004). On the

[5]Corresponding author.
E-mail shantanuc@igib.res.in; fax 91-11-2766-7471.
Article is online at http://www.genome.org/cgi/doi/10.1101/gr.4508806.

16:644–655 ©2006 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/06; www.genome.org

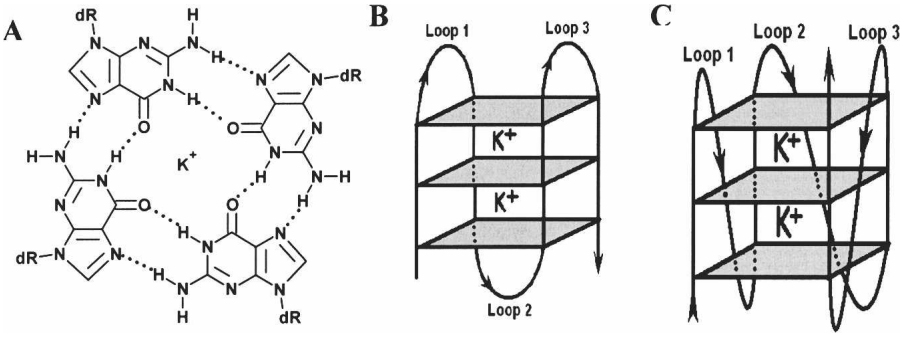

Figure 1. Schematic representation of G4 motif. (*A*) Hydrogen-bonded G-tetrad with K^+ (Na^+ also stabilizes a G-tetrad); each guanine in this planar array is contributed from different G-runs, which are separated by intervening loops in an intramolecular motif. (*B,C*) Intramolecular folding pattern showing stem and loop organization in an antiparallel (*B*) and parallel (*C*) conformation of a G4 motif, where the planes represent each tetrad unit and are stacked to form the stem of the motif.

other hand, non-B DNA forms have been implicated as regulatory signals in *E. coli* under supercoiling stress. Specific roles have been illustrated in a few cases like the *ilvGMEDA*, *leuV*, and *ilvYC* operons (Sheridan et al. 1999; Opel and Hatfield 2001; for review, see Hatfield and Benham 2002). In this context, it is interesting to consider that G4 DNA might be important in gene regulation and genetic stability in prokaryotes.

Using a nucleic acid pattern recognition program, we searched 18 representative prokaryote genomes for G4 DNA sequences and analyzed their genomic distribution and association with genes. Our analysis indicated enrichment of G4 DNA within the near upstream region of genes relative to other non-coding regions across all organisms. A comparative functional analysis (using 23 classes from COGS) of >61,000 open reading frames (ORFs) indicated that transcription, amino acid biosynthesis, and signal transduction genes could be predominantly controlled by G4 DNA. We also observed that the motifs were conserved within promoters of orthologous genes across phylogenetically distant organisms. Additionally, randomly selected potential G4 forming sequences from *E. coli* were observed to adopt quadruplex structure in solution under physiological conditions. Transcription-factor-binding site analysis of 55 DNA-binding proteins in the region flanking G4 DNA sequences in *E. coli* indicated significant association with global regulators, which are known to be supercoiling sensitive. Taken together, our findings indicate a putative role of G4 DNA in prokaryotic gene regulation. Based on our observations in *E. coli*, we predict that G4 DNA may be one of the factors involved in DNA-topology-mediated gene expression.

Results

Definition of G4 motifs, classification, and genome-wide search strategy

Intramolecular G4 DNA motifs comprise four runs of guanines (constituting the stem of G4 motif) interspersed with nucleotide bases, which form three intervening loops (Fig. 1; Balagurumoorthy and Brahmachari 1994; Gilbert and Feigon 1999). We developed a pattern search algorithm to identify potential G4 DNA sequences wherein four consecutive G-runs were identified, after allowing for three intervening loops (see Methods). In order to avoid overestimation of G4 DNA motifs, overlapping patterns (with more than four G-runs) were stitched together and the sequence was designated as a tract, which can adopt multiple G4 motifs but is most likely to present only one exclusive motif. In the following text, we refer to such tracts as PG4 (potential G4) motifs. Applying our search strategy in a genome-wide screen, we collated two basic forms of information for mapping and comparative analyses: (1) the frequency of the bases comprising the tracts and (2) association of the tracts with the regulatory regions of genes.

Results of genome searches

We applied our search strategy to 18 complete prokaryote genomes representing different phylogenetic origins. All PG4 motifs identified within the respective genomic regions—intragenic, putative regulatory (up to 200 bp upstream of genes), or "rest-of-intergenic" (see Methods)—for 18 organisms are listed, organized according to the above criteria, on our Web site (http://www.igib.res.in/prokaryote/PG4.htm). Table 1 shows a summary of the distribution in both + and − strands.

The overall number of motifs was similarly distributed in both the strands and appeared to be higher in organisms with a high GC%, which was expected as the motifs were G-rich (Table 1; Supplemental Table S1). Interestingly, the frequency of PG4 motifs (number of bases involved in motif formation per kilobase) varied considerably among the three regions. It was interesting to observe that >98% of the motifs in both the + and − strands had two tetrad units in the stem, and a tract size below 40 bases was prevalent (>95%) across all organisms in both strands (Supplemental Tables S3 and S4). The variation in size of the three loops was also analyzed and is represented in mosaic plots (Supplemental Fig. S2; Friendly 1994) for each genomic region across all organisms. The overall distribution indicates no preference in any particular loop size combination in the intergenic regions (plots A and B), while the intragenic region (plot C) showed a preference for a loop size of four in all combinations. All genomes were enriched in PG4 motifs vis-à-vis the probability of random occurrence. By using BLAST for each identified PG4 motif sequence against the respective organism, we observed that the probability of random occurrence was very low for most sequences except ones that were <14 bases and occurred multiple times (Supplemental Table S2) (an extensive set from five different organisms is available at http://www.igib.res.in/prokaryote/PG4.htm). Several other independent observations emphasize this (see Supplemental material).

Distinct genomic distribution was observed for PG4 motifs

The frequency of motifs (both strands together) in the different genomic regions is listed in Table 1. *Streptomyces coelicolor* A3(2) (Sco), with the highest GC% (72.11%) (Supplemental Table S1) had the highest density of PG4 motifs, and the near upstream region harbored the major proportion of them. Similarly, *Pseudomonas aeruginosa* PAO1, *Halobacterium* sp. NRC-1, and *Xanthomonas campestris* pv. *Campestris* str. ATCC 33,913 also showed a high PG4 motif frequency in the −200-bp region. On the other hand, the low-GC-content (<40%) genomes of *Clostridium acetobutylicum* ATCC 824, *Lactococcus lactis* subsp. *lactis* Il1403, *Haemophilus influenzae* Rd KW20, and *Mycoplasma genitalium* G-37 had a low frequency of motifs. Figure 2A shows the distribution of PG4 motifs in intragenic, intergenic (beyond −200 bp), and

Table 1. Overall distribution of PG4 motifs in 18 prokaryotes

	Intragenic			Regulatory (−200 bp)			Rest-of-intergenic		
		Number of PG4 tracts[b]			Number of PG4 tracts[b]			Number of PG4 tracts[b]	
Organism	Frequency[a]	+ strand	− strand	Frequency[a]	+ strand	− strand	Frequency[a]	+ strand	− strand
Afu	40.0	2232	2126	23.0	74	77	9.9	41	70
Bsu	12.3	1303	1243	10.0	114	120	6.2	155	125
Bba	29.2	2771	2842	21.1	120	146	37.9	69	62
Cac	4.0	383	415	2.7	37	23	1.2	47	18
Eco	27.0	3029	3148	11.7	138	113	9.0	177	149
Hin	6.4	279	275	19.2	16	16	8.5	57	36
Hal	83.9	4037	4024	97.9	403	387	57.4	250	255
Lla	4.7	275	275	1.1	4	11	1.5	17	31
Mth	42.7	1873	1832	25.1	90	72	9.3	100	28
Mbo	113.8	10,462	10,909	100.4	704	791	77.3	436	385
Mtu	113.3	10,614	10,971	99.9	703	797	74.1	476	450
Mge	3.3	49	45	3.1	2	2	2.0	26	27
Pae	116.4	16,055	17,376	123.8	1422	1560	93.4	725	871
Pfu	14.2	714	682	26.3	95	79	16.1	58	41
Stm	33.3	3910	3927	12.9	160	149	9.2	234	230
Sco	198.0	36,539	35,720	230.0	3662	3509	168.5	2297	2201
Vch	14.5	1100	1057	7.5	54	40	4.6	97	89
Xcc	88.5	9961	9937	104.0	1006	993	44.7	833	841

Species acronyms from KEGG, mentioned in Methods (http://www.genome.jp/kegg/catalog/org_list.html).
[a]Number of bases contributing to PG4 motifs per kilobase of sequence in + strand.
[b]Number of sequence elements with possibility of multiple G4 DNA conformations.

putative regulatory regions for the + strand in 18 bacterial genomes. Owing to the observed overall correlation between motif density and GC% in the respective region, as expected for G-rich elements, we analyzed the variation in motif density across the three genomic regions after normalizing for GC%. Here, we excluded the downstream intergenic regions between convergent genes to avoid PG4 motifs, which could be putative terminators—as indicated by G4 DNA-induced polymerase "falloff" in several studies (Simonsson et al. 1998; Siddiqui-Jain et al. 2002). These regions have been independently analyzed relative to regulatory regions (see below). PG4 motifs in each region were expressed as a ratio of the frequency of GC bases in the respective region ($R_{PG4/GC}$) (Fig. 2A). A higher R was observed in the intragenic regions relative to the intergenic region (beyond −200 bp) in the + strand ($P < 0.001$; $\text{median}_{\text{intragenic}} = 0.030$, $\text{median}_{\text{intergenic (beyond } -200 \text{ bp)}} = 0.014$) (Fig. 2A, inset). More interestingly, the frequency of PG4 motifs in the putative regulatory region (up to −200 bp) was observed to be higher in comparison to the intergenic region ($P < 0.0025$; $\text{median}_{-200 \text{ bp}} = 0.025$) (Fig. 2A, inset; Supplemental Table S5). The observed difference was true for both the + and the − strands independently (Supplemental Fig. S3 shows − strand distribution); however, the difference in distribution of PG4 motifs between the strands was not statistically significant in any region (Supplemental Table S5). Sequence with multiple G-runs that would not adopt a G4 DNA motif (control pattern) did not show enrichment in the putative regulatory regions (Supplemental Fig. S9; Supplemental Table S12). We also checked the distribution of PG4 motifs using the variable stem size parsing method (Huppert and Balasubramanian 2005; Todd et al. 2005) and found it to be consistent with our observations (Supplemental Fig. S8).

Motif frequency decreases beyond 50–100 bases upstream of start codon

We tested the implications of the above observation with respect to the near upstream region by plotting the number of bases making PG4 motifs in the + strand within blocks of 50 bases up to 500 bases upstream of all genes, excluding coding regions (Supplemental Fig. S4). The motif frequency decreased sharply on moving upstream from the start codon in nearly all organisms, indicating a prevalence of PG4 motifs in near upstream regions. We checked whether the GC% of the entire intergenic region affected the motif frequency in the near upstream (−100 bp) region. The ratio of $R_{PG4/GC}$ in the −100-bp region and $R_{PG4/GC}$ in the entire non-coding region excluding the first −100 bases from a gene ($R_{PG4/GC(-100 \text{ bp})}/R_{PG4/GC(\text{beyond } -100 \text{ bp})}$) was plotted against the overall GC% of the entire intergenic region of all 18 organisms (Fig. 2B). Interestingly, a positive correlation was observed in this case (Spearman's ρ [nonparametric correlation coefficient] = 0.72, $P < 0.001$). As a control, we analyzed the frequency ratios in the entire intergenic versus intragenic region ($R_{PG4/GC(\text{intergenic})}/R_{PG4/GC(\text{intragenic})}$) against the overall genomic GC%. In this case, the correlation was not significant (Spearman's $\rho = 0.09$, $P = 0.499$) (Fig. 2C). This indicated that GC-rich genomes positively selected for PG4 motifs in the near upstream region relative to respective far intergenic regions. A particular case is the GC-rich *S. coelicolor* genome, where we clearly observed a higher occurrence of PG4 motifs in the putative regulatory region relative to the other genomic regions (Fig. 2A) and also relative to other genomes (Supplemental Fig. S4). Overall distribution of motifs flanking 0.5 kb of the start codon of genes (in 61,355 ORFs) across 18 organisms (Fig. 2D) highlighted the decrease in PG4 frequency (on moving away from start codon) in the near upstream region vis-à-vis the coding region and also indicated that sequence overlapping start codons was relatively less dense in motifs. Interestingly, intragenic regions close to the start codon also showed high PG4 motif density, which could be potential signals for repression of transcription. This is in line with observations of G4-motif-induced arrest in DNA synthesis (Woodford et al. 1994; Simonsson et al. 1998; Siddiqui-Jain et al. 2002). Additionally, the possibility of attenuation or antitermination signals by G4 motif formation in the transcribed mRNA

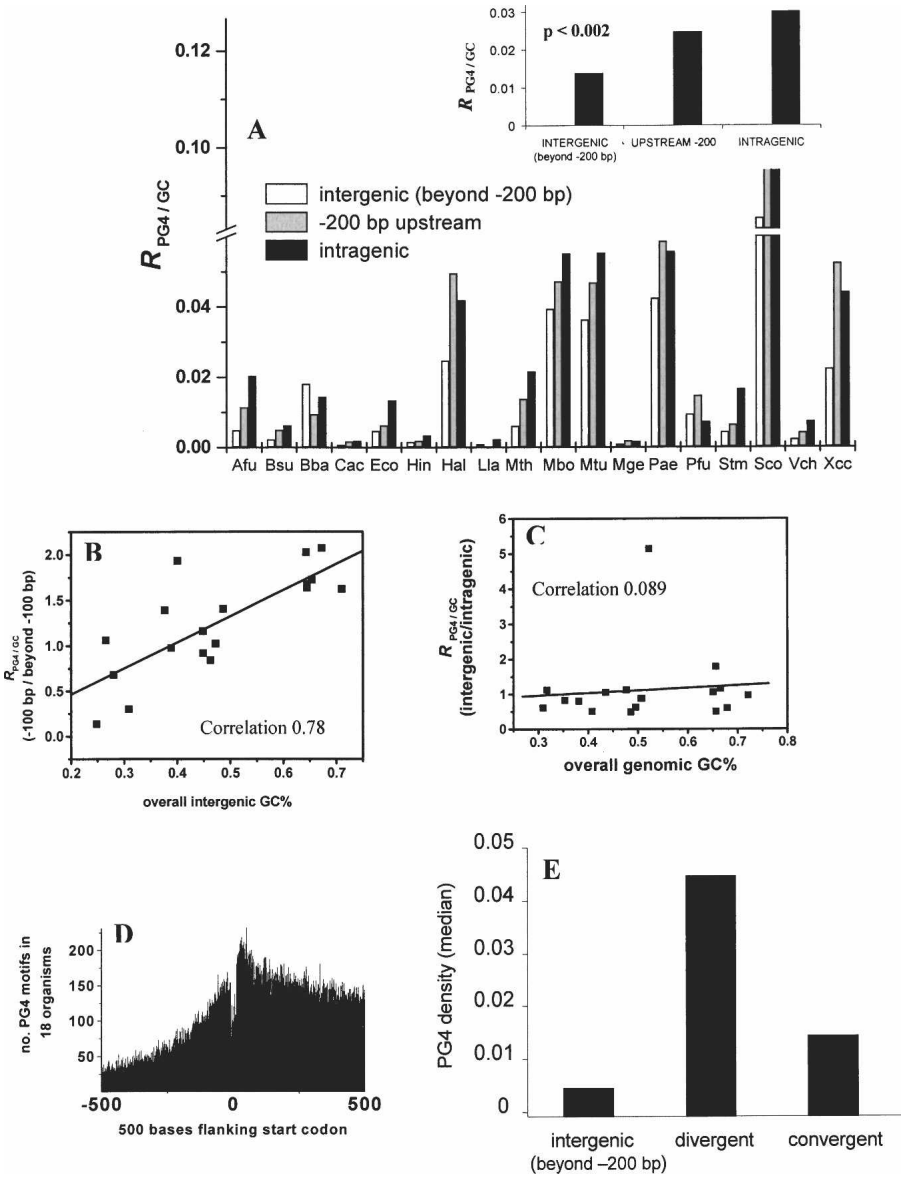

Figure 2. Putative regulatory regions in prokaryotes are enriched in PG4 motifs. (*A*) Genome-wide distribution of PG4 motifs within the + strand in 18 prokaryotes showing frequency of the bases forming PG4 motifs in each region expressed as a ratio of the GC frequency of the respective region ($R_{PG4/GC}$) for each organism. (*Inset*) Median ratio ($R_{PG4/GC}$) for each region calculated from the distribution in the respective regions across all organisms. (Supplemental Table S5 shows the mean and standard deviation, and Supplemental Fig. S3 shows a similar distribution for the − strand.) The intergenic (beyond −200 bp) region includes all intergenic regions except the downstream region between two convergently oriented genes. (*B*) GC-rich organisms have selected for PG4 motifs in their immediate upstream regions. Ratio of the frequency of PG4 motifs (after controlling for GC% in the respective regions) in the −100-bp region versus beyond −100 bp within the intergenic region shows a high correlation with the GC% of the intergenic region for respective organisms. (*C*) The motif frequency of intergenic versus intragenic regions does not depend on the GC% of the genome. The ratio-plot for intergenic versus intragenic regions against overall (genome-wide) GC% of the organism shows very low correlation. *M. genitalium* shows a high ratio (>5.0) because of a very low intergenic basepair length (correlation on excluding *M. genitalium* was 0.24). (*D*) The number of PG4 motifs decreases sharply on moving upstream of genes relative to the intragenic regions. Data were plotted from all 61,355 ORFs in 18 organisms within the flanking 500 bases of the start codon of all ORFs. The center of each motif sequence was used for mapping with respect to the start codon (i.e., for a sequence of length *n*, the *n*/2-th base was used as its coordinate). (*E*) Promoter-rich regions have a higher density of PG4 motifs. Intergenic regions separating divergently (promoter-rich) and convergently (possibly promoter-less) oriented gene pairs were mapped in all 18 organisms for comparison. The median of PG4 density (number of bases involved in motif pattern normalized for sequence length of the respective region) is shown along with the density in the intergenic regions (beyond −200 bp, as in *A*). The difference between the divergent and convergent ($P < 0.007$) and the divergent and intergenic ($P < 0.025$) regions was significant, while the difference between the convergent and intergenic regions was not significant ($P = 0.199$). All statistical comparisons were done in a pairwise mode for the different genomic regions, and significance was estimated using the two-tailed nonparametric Signed Wilcoxon Test. The organism acronyms are as obtained from KEGG and are mentioned in Methods.

cannot be ruled out (Christiansen et al. 1994; Horsburgh et al. 1996).

Intergenic regions separating divergently oriented genes are enriched in PG4 motifs

We analyzed intergenic regions separating divergently and convergently oriented genes in 18 organisms for PG4 motif density. It was observed that the motif density was significantly higher in the divergent intergenic regions ($median_{divergent}$: 0.0469 vs. $median_{convergent}$: 0.0158; $P < 0.007$) (Fig. 2E). We also observed that PG4 density in divergent intergenic regions was higher than the density observed in the intergenic region ($median_{intergenic(beyond\ -200\ bp)}$: 0.0055; $P < 0.025$). Although a higher PG4 density was observed in divergent intergenic relative to putative regulatory regions, this was not statistically significant ($P = 0.199$) (Supplemental Table S10). Similarly, the difference with intragenic regions was also not significant. PG4 density in convergent intergenic regions, however, did not show significant difference when compared to intergenic, intragenic, or regulatory regions (Supplemental Table S10). Thus, although a functional role of PG4 motifs as terminators cannot be ruled out, enrichment in divergent intergenic regions relative to convergent ones suggests a functional role of PG4 motifs with regulatory consequences.

ORFs with PG4 motifs in regulatory region show distinct functional distribution

We analyzed 37,974 ORFs across 18 species in 23 different functional classes from the COGS database (Tatusov et al. 1997). These ORFs were considered after excluding genes belonging to undefined functional classes (i.e., function unknown and general function prediction only). Of these, 5574 (14.7%) ORFs had at least one motif in the + or the − strand within the −200 bp of the start codon. We observed that the functional classes secondary metabolite biosynthesis, transport and catabolism (25.52%), transcription (25.64%), and signal transduction (24.08%) had more genes, which harbored one or more PG4 motifs in their regulatory regions relative to others ($P < 0.004$; the average genes in other classes is 11.97%, SD = 3.75%) (Fig. 3A). Interestingly, >17% (968 genes) of all ORFs having motifs in the regulatory region pertained to transcription ($P < 10^{-8}$; <10% in any other class) (Supplemental Fig. S5). We did not observe much variation in the intragenic PG4 motif frequency between the function classes (~15–23 bases per kilobase of gene) on analyzing the coding region of all 37,974 ORFs (Fig. 3B). Transcription factor genes showed a somewhat higher motif frequency than the average; however, this was not significant ($P = 0.108$). It must be mentioned that regulatory control, considering operon organization within bacterial genomes, may not be necessarily exerted from the immediate upstream region of the ORFs, and thus our analysis gives a global view of the PG4 motifs vis-à-vis their putative functional role. Analysis in the context of operons in *E. coli* is presented below.

Orthologs in distantly related species conserve PG4 motifs within regulatory region

Conservation of motifs across species, especially if the species belong to evolutionarily distant groups, indicates biological significance as functional signals. We hypothesized that if PG4 elements serve as regulatory motifs, they are liable to be conserved in the regulatory region of orthologous genes. We analyzed orthologous groups from COGS for *E. coli* genes harboring PG4 motifs in the –200 region and checked whether the corresponding orthologs (in the 17 other organisms) also had one or more PG4 motifs in their upstream region. We found 40 genes where PG4 motifs were conserved (Table 2). In 36 of these, at least one species was from an evolutionarily distant group (von Mering et al. 2003), and 20 genes showed conservation in orthologs across at least two distant groups. It was interesting to observe that a majority of the genes with conserved motifs pertained to metabolism (amino acid, carbohydrate, and vitamins/cofactors), membrane transport (ABC and ion), transcription, and translation.

Analysis of PG4 motifs in *E. coli*

Mapping of PG4 motifs to the regulatory network in E. coli

Based on our findings, which collectively indicated that PG4 motifs may be biologically significant as functional regulators, we focused on *E. coli* as a reference organism for mapping the motifs to known regulatory networks. *E. coli* was chosen for this analysis as by far it is the most studied organism in this respect. We mapped all the PG4 motifs found within the upstream region (200 bases) on the + strand of genes in *E. coli* in the context of characterized/predicted promoters and operons (Salgado et al. 2004) (see Supplemental material). Operons with PG4 motifs in the regulatory region are listed in Table 3.

Target sites for global supercoiling–sensitive regulators are predominantly associated with PG4 motifs in E. coli

We mapped transcription-factor-binding sites (TFBS) to sequence flanking (100 bases) PG4 motifs, which were present within −200 bases of the start codon. The target sites of nine DNA-binding proteins (Crp, FIS, GlpR, Lrp, OmpR, RopD, RpoS, SoxS, and TyrR) were prominent (with >50 sites associated with 118 PG4 motifs in the + strand, for each factor). RpoD (38.8%) and Lrp (34.7%) constituted the majority of 6493 sites observed for 28 factors on the + strand (Fig. 4A). Similarly, out of 4250 sites for 26 factors associated with 96 PG4 motifs present on the − strand, RpoD and Lrp comprised 38.9% and 33.9% of the sites. FIS, GlpR, and RpoS constituted 4%–6%, and Crp, OmpR, TyrR, and SoxS had ~1%–2% sites each associated with PG4 motifs present on the + or − strand. As a control set, 445 putative promoter regions (up to 200 bases upstream of start codons) devoid of PG4 motifs were also considered for TFBS. Out of a total of 14,292 TFBS in this case, RpoD and Lrp constituted 34.8% and 37.1% of the sites, respectively. The frequency of each target site in the three sets—that is, sequence flanking PG4 motifs present on the + strand, flanking regions of motifs present on the − strand, and control promoter regions with no motifs—was compared. Interestingly, a significant ($P < 0.001$; nonparametric comparison using the Mann-Whitney Test) difference in the frequency of individual sites was observed in the flanking region of PG4 motifs (both, when present on the + or − strand) with respect to the control set, for five of the nine factors considered: Lrp, FIS, GlpR, RpoS, and RpoD. The frequency distribution of target sites was not different when associated with PG4 motifs present on the + or − strand ($P > 0.05$), except in the case of GlpR. The *P*-values for all comparisons are given in Supplemental Table S8. Individual frequency distribution plots for target sites of the nine factors in the respective regions and average sites (median) per motif (or pro-

Table 2. Representative table showing *Escherichia coli* genes and orthologs with conserved PG4 motifs (within regulatory region)

	Gene name	*E. coli* ORF	PG4 position from start codon	Organisms with conserved PG4 motif in orthologs[a]	Function[b]	Biological process[b]
1.	*MalK*	b4035	−94	Stm; Sco	ABC transporter	Environmental information processing
2.	*FepB*	b0592	−60	Hal; Mbo; Mtb; Pae; Sco; Vch	ABC transporter	Environmental information processing
3.	*GabP*	b2663	−118	Bsu; Mbo; Mtb; Pae; Stm; Xcc	Ion coupled membrane transporter	Environmental information processing
4.	*YhiP*	b3496	−25	Stm; Sco; Xcc	Ion coupled membrane transporter	Environmental information processing
5.	*CutA*	b4137	−52	Xcc	Inorganic ion transport	Environmental information processing
6.	*RcsC*	b2218	−79	Bsu; Pae	Two-component signal transduction	Environmental information processing
7.	*Ffh*	b2610	−13	Pae; Stm; Xcc	Protein transporter, secretion	Genetic information processing
8.	*UvrC*	b1913	−14	Hal; Sco	Replication, recombination, and repair (repair of UV damage to DNA)	Genetic information processing
9.	*HepA*	b0059	−77	Stm	Replication, recombination, and repair (transcriptional activator)	Genetic information processing
10.	*PriA*	b3935	−149	Stm	Replication, recombination, and repair (putative helicase)	Genetic information processing
11.	*OxyR*	b3961	−18	Pae; Xcc	HTH family transcriptional regulator (activator, hydrogen peroxide-inducible genes)	Genetic information processing
12.	*RpsA*	b0911	−45	Stm; Sco	Structural constituent of ribosome	Translation
13.	*RpmE*	b3936	−54	Mbo; Mtb; Stm; Sco	Structural constituent of ribosome	Translation
14.	*TufB*	b3980	−157	Stm; Sco; Vch; Xcc; Hin	Protein chain elongation factor, responds to osmotic stress	Translation
15.	*QueA*	b0405	−72	Pae	tRNA synthesis	Translation

The full list is given in Supplemental Table S6.
[a]Species acronyms as in KEGG (http://www.genome.jp/kegg/catalog/org_list.html). Groups of phylogenetically related organisms were obtained from the STRING server (see Methods).
[b]Annotation information from COGS, the GO server, and NCBI.

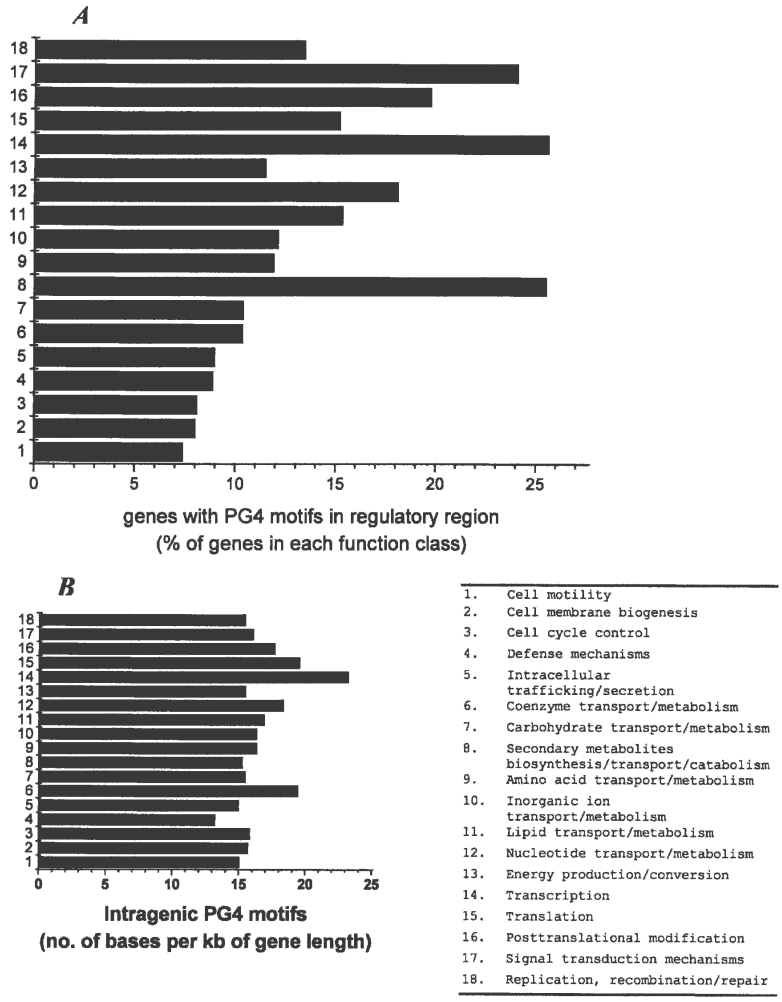

Figure 3. Genes harboring PG4 motifs in their regulatory region show distinct functional distribution in a comparative analysis comprising 37,974 ORFs from 18 organisms. (*A*) The distribution of genes with at least one PG4 motif within the −200-bp region is shown as the percentage of total genes in the respective function class—secondary metabolite biosynthesis, transcription, and translation related genes show significant difference ($P < 0.004$). (*B*) The intragenic PG4 motif density indicates that the distribution is not significantly different across the functional classes ($P = 0.108$). The PG4 motif density was calculated as the number of bases involved in motif formation per kilobase of gene length. Two classes, chromatin structure and dynamics and RNA processing and modification, which constitute only 0.054% and 0.09% of the distribution, were not included in the plots. Extracellular structure, nuclear structure, and cytoskeleton genes do not have any motifs in their regulatory regions. Undefined classes like function unknown and general function prediction have been excluded from analysis along with genes not found in the COGS database. All function information was obtained from the COGS database. A plot showing distribution across the functional classes with respect to the total ORFs (5574) with PG4 motifs in −200-bp regions is shown in Supplemental Figure S5.

moter, in case of the control set) for the five factors with significant difference are shown in Figure 4, B and C (also Supplemental Fig. S6). We used predicted factor-binding sites as observed before without any further change. A large number of binding sites for factors like RpoD and Lrp result from the presence of numerous contiguous sites (overlapping within 2–3 bases at times); however, this is not expected to affect our conclusions since it holds for the control set also.

E. coli *PG4 sequences adopt G–quadruplex motifs in solution*

All PG4 motifs identified by us were based on previous information about sequence patterns, which could adopt quadruplex motifs. We selected 11 sequences randomly from the upstream region of different genes in *E. coli* and checked their potential to adopt a quadruplex motif in solution under physiological conditions using CD. CD profiles for both strand orientations of the quadruplex motifs, parallel and antiparallel, have been well established (Balagurumoorthy and Brahmachari 1994). Nine out of 11 sequences readily formed structure in the presence of monovalent cation (both Na^+ and K^+). The CD spectra for all 11 sequences are included in Supplemental Figure S7.

Discussion

Our analysis shows overrepresentation of G-quadruplex or G4 DNA motifs in putative regulatory regions in the noncoding genome of prokaryotes (Fig. 2A–E). Interestingly, a detailed analysis in regulatory regions of *E. coli* indicated that the target sites of transcription factors Lrp, FIS, GlpR, RpoS, and RpoD were predominantly associated with G4 motifs. This is the first genome-wide comparative study of G4 DNA in prokaryotes, and collectively our observations suggest that PG4 motifs may be biologically relevant as regulatory signals in prokaryotes. This is further supported by the fact that genomes with high PG4 motif frequency in their regulatory region (Fig. 2A) also show strong emphasis on regulation [12.3% and 8.4% of the coded proteins in *S. coelicolor* A3(2) and *P. aeruginosa* PAO1, respectively, are predicted to be involved in regulation relative to 3.0% in *Mycobacterium tuberculosis*, 5.8% in *E. coli*, and 5.3% in *Bacillus subtilis*] (Stover et al. 2000; Bentley et al. 2002).

Recent studies show evidence of in vivo presence of G4 DNA in *E. coli* (Duquette et al. 2004) and prevalence of G4 DNA in the human genome, which indicated a possible functional role of these motifs (Huppert and Balasubramanian 2005; Todd et al. 2005). In comparison to the motifs reported from the human genome, the loop size distribution in prokaryotes appears quite contrasting. Our analysis indicated loop sizes of >3 nt to be predominant (Supplemental Fig. S2) compared to an overrepresentation of single nucleotide loops in the human genome (Huppert and Balasubramanian 2005). Based on analysis of the CD spectra of several G-rich oligonucleotides (Dapic et al. 2003; Hazel et al. 2004), this indicates the likelihood of G4 motifs with parallel strand orientation (Fig. 1) being preferred in the human genome, while the bacterial genomes appear to predominantly harbor motifs that could adopt both parallel and antiparallel structures (Supplemental Fig. S7).

Table 3. Representative table showing *Escherichia coli* operons with PG4 motifs in the putative regulatory region (up to −200 bp upstream of operon) for the + strand

Operon (RegulonDB) [a]	PG4 position from start codon	First gene in operon	Function of the first gene (COGS)
YadE	−17	yadI b0129	Carbohydrate transport and metabolism genes
YkfG_yafX	−184	yafW b0246	Not in COGS
afuC_b0263_insB_2_insA_2	−40	yagB b0266	Not in COGS
YagA	−131	yagA b0267	Replication, recombination, and repair genes
yagE_yagF	−143	yagE b0268	Amino acid transport and metabolism genes
queA_tgt_yajC_secD_secF	−72	queA b0405	Translation genes
Apt	−89	apt b0469	Nucleotide transport and metabolism genes
YbbB	−187	allS b0504	Transcription genes
YbbT	−43	allA b0505	Nucleotide transport and metabolism genes
ybhR_ybhS_ybhF_b0795_ybiH	−188	ybiH b0796	Transcription genes
YliH	−24	yliH b0836	Translation genes
b1180	−37	ycgL b1179	Not in COGS
DsbB	−63	nhaB b1186	Inorganic ion transport and metabolism genes
Kch	−95	cls b1249	Lipid transport and metabolism genes
CybB	−95	cybB b1418	Energy production and conversion genes

The full list is given in Supplemental Table S7.
[a]Predicted or characterized operon from RegulonDB.

Growth phase response of sigma factor RpoD (σ^{70}) and global regulators FIS and Lrp may be mediated through G4 motif formation in their target sites during supercoiling stress

Transcription-factor-binding site analysis in *E. coli* indicated five regulators, Lrp, FIS, GlpR, including the sigma factors σ^{70} and σ^{S} (product of *rpoD* and *rpoS* genes, respectively), to be predominantly associated with PG4 motifs. Except σ^{S}, which is essential for transcription of stationary-phase genes, σ^{70}, Lrp, FIS, and GlpR are "switched on" in the exponential growth phase following nutritional upshift and together control >1000 genes (Ishihama 1999 and references therein; Martinez-Antonio and Collado-Vides 2003). It is interesting to consider the relevance of association with PG4 motifs for these crucial regulatory factors.

An increase in the ATP/ADP ratio (energy charge) due to nutritional upshift enhances gyrase activity (because of the enhanced phosphorylation potential of the cell) resulting in higher negative superhelicity (Balke and Gralla 1987; Travers et al. 2001; Hatfield and Benham 2002). The stress-induced duplex destabilization (SIDD) model of Benham et al. indicates that high superhelical density results in the formation of localized non-B DNA motifs to counter superhelical stress, which exert regulatory control in *E. coli* (Singh et al. 1995; Hatfield and Benham 2002). In line with the SIDD model, we envisage that occurrence of local sites of partially destabilized duplex states may induce strand- and sequence-specific formation of G4 motifs on strand separation (Wang et al. 2004). Based on our finding that regulatory regions harboring G4 sites are strongly associated with target sites of FIS, Lrp, GlpR, σ^{70}, and σ^{S} (Fig. 4), we predict that the regulatory response of these factors is mediated by G4 DNA formation in the supercoiled state. The actual mechanism, however, may involve presentation of specific recognition elements along with a combinatorial effect of diverse factors ranging from DNA topology to cellular interacting partners. The association of the target sites over a broad region (100 bp flanking G4 motifs) is consistent with the observation that transcription factor binding induces transmission of superhelical destabilization to distal promoter sites (Sheridan et al. 1999; Opel et al. 2004).

Our prediction is also supported by a high incidence of in vivo G4 motifs in supercoiled DNA relative to other topoisomers (58% vs. 31% and 24% in relaxed and linear templates, respectively) in *E. coli* (Duquette et al. 2004). The association of FIS-binding sites with G4 motifs is consistent with several reports indicating that FIS operates by stabilizing local DNA architecture and that supercoiling-responsive promoters harbor FIS-binding sites (Schneider et al. 2000). Additionally, it has been demonstrated that *topA* induces G4 motif formation (Arimondo et al. 2000). This is in line with our prediction, as FIS has been observed to induce *topA* expression during oxidative stress (Weinstein-Fischer et al. 2000).

Target sites for σ^{S}, which regulates >100 stationary-phase-specific genes (Ishihama 1999; Martinez-Antonio and Collado-Vides 2003), were also associated with PG4 motifs. Decreased DNA superhelicity in stationary-phase *E. coli* cells induces σ^{S} activity while repressing that of σ^{70} (Kusano et al. 1996; Bordes et al. 2003). On the other hand, σ^{S} response can also be induced by high osmolarity and heat shock during growth phase (Hengge-Aronis 1993; Jishage et al. 1996). Thus, although σ^{S} response is complex, it is tempting to speculate that G4 motif formation in growth phase may assist repression of σ^{S} target sites and thereby contribute in switching of the σ^{70}/σ^{S} response during the transition from growth to stationary phase (Ishihama 1999).

A recent study demonstrates that the osmotic stress response (OSR) in *E. coli* is induced by supercoiling. Interestingly, they observed that the upstream region of genes in OSR are significantly enriched in binding sites for FIS, Lrp, GlpR, and RpoS (Table 5 of Cheung et al. 2003), which is consistent with our prediction. We also observed that several genes with statistically significant expression in OSR, like *apt, poxB, nhaB, ydcF, ygfF*, and *yibK*, had one or more PG4 motifs in their regulatory regions. In another recent genome-wide expression study of genes responsive to DNA relaxation in *E. coli*, we observed that several significantly induced (*cls, ycgL, insB_2*, and *insA_2*) and repressed (*eaeH, yeeS, mazG*, and *gidA*) genes may be potentially regulated by PG4 motifs, present either in the respective promoters or upstream of respective operons (Peter et al. 2004).

It must be noted that our conclusions are based on mostly predicted transcription-factor-binding sites and unlike the human *c-myc* and other cases (where G4 is implicated to be regulatory) (Howell et al. 1996; Siddiqui-Jain et al. 2002; Etzioni et al. 2005), no experimental proof exists of G4 DNA-mediated regulation in bacteria. Genome-wide ChIP analysis for bind-

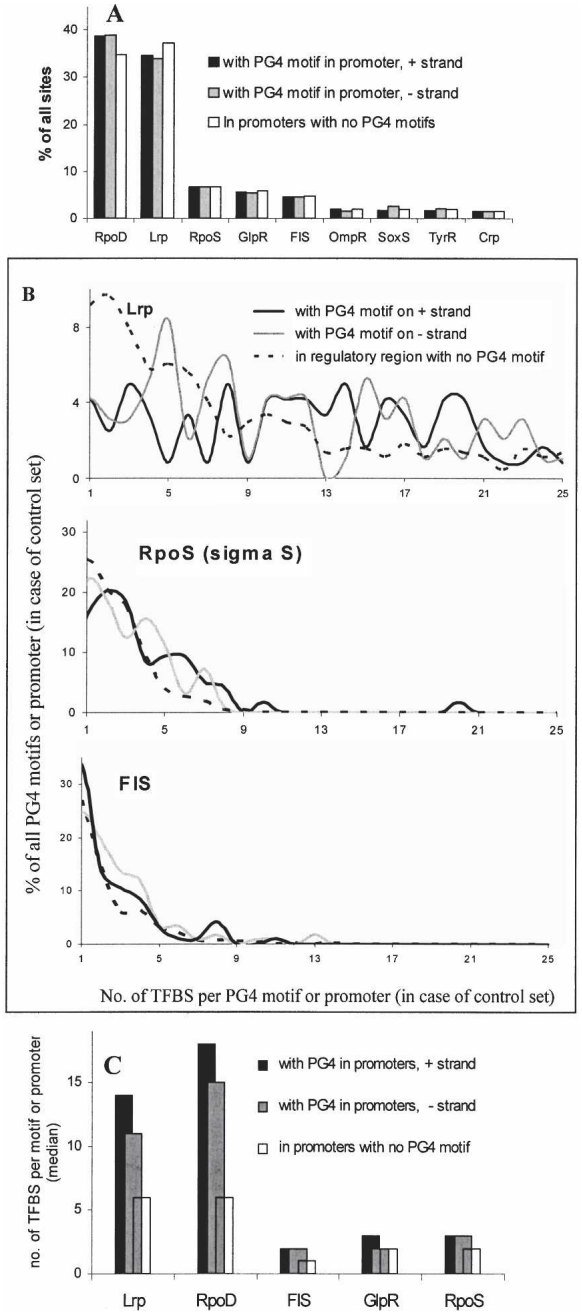

Figure 4. Global regulators Lrp, FIS, and GlpR and sigma factors σ^{70} and σ^{S} are predominantly associated with PG4 motifs in *Escherichia coli*. We computationally mapped target sites for 55 DNA-binding proteins in the region flanking (100 bp) PG4 motifs present within −200 bp of start codons in the + strand (118 motifs) and − strand (96 motifs). Sites were also mapped to 445 promoter regions (within −200 bp of start codon) devoid of PG4 motifs as a control set. (*A*) Overall representation of sites (for nine factors with >1% sites) as a percentage of total sites for 55 DNA-binding proteins is shown for the respective regions. (*B*) Frequency distribution of TFBS. Motifs or promoters (%) were plotted against the number of sites found either flanking the motifs or within the promoter (in case of control set); representative plots for three factors are shown (for others, see Supplemental Fig. S6). Distributions were observed to be significantly ($P < 0.001$) different for Lrp, RpoD, FIS, RpoS, and GlpR when compared between the + or − strand and the control set, while SoxS, TyrR, Crp, and OmpR did not show a statistically significant difference ($P > 0.05$). (*C*) Target sites (median) per motif (+/− strand) or promoter (control set) are shown for five factors with significantly different distribution. Nonparametric comparisons were done using the Mann-Whitney U-test; the *P*-values for respective comparisons are shown in Supplemental Table S8.

ing sites in conjunction with molecules or factors that specifically bind to G4 motifs and in vitro specificity assays will be required to confirm our findings (Schaffitzel et al. 2001; Rezler et al. 2005).

Our findings provide several bioinformatics insights. Significant among them are tentative results showing that transcription, secondary metabolite biosynthesis, and signal transduction classes have more genes (relative to other classes) that could be under G4 motif control (Fig. 3A). This is largely consistent with the fact that supercoiling-induced DNA topology controls transcription of several regulons and stimulons during the exponential growth phase (Schneider et al. 1999; Hatfield and Benham 2002; Peter et al. 2004). Cheung et al. (2003) have made similar observations with respect to clusters of genes belonging to macromolecule, amino acid, and building block biosynthesis, which are significantly overexpressed as a result of superhelical stress during the osmotic stress response (OSR). Another interesting observation indicated that GC-rich genomes selected positively for G4 motifs in the near upstream regions (Fig. 2B). Based on the SIDD model of Hatfield and Benham, it is tempting to speculate that in GC-rich organisms, where duplex destabilization is energetically more demanding, favorable G4 motif formation may relax superhelical density (Hatfield and Benham 2002). Our previous reports suggesting favorable kinetics of G4 motif formation within the nuclease-hypersensitive element of the human *c-myc* promoter support this hypothesis (Halder and Chowdhury 2005; Halder et al. 2005). However, such a possibility has to be clearly demonstrated.

Taken together, these findings suggest that G4 DNA may be biologically relevant as regulatory signals in prokaryotes. The motifs may be particularly important in translating environmental stimuli (nutritional upshift) into a functional message by presenting target sites for orchestrating the activity of global transcriptional regulators in *E. coli*.

Methods

Organisms

18 bacterial organisms were used for our analysis after downloading their genomes from the NCBI database (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi). (The abbreviations used are from the KEGG database; http://www.genome.jp/kegg/catalog/org_list.html; Kanehisa and Goto 2000; they are also mentioned in the Supplemental material). Organisms were chosen such that apart from two groups of closely related organisms—Vch, Stm, Eco and Mtu, Mbo, Sco—all others belonged to evolutionarily distant clades. The STRING server (http://string.embl.de/; von Mering et al. 2003), which is based on various phylogenetic distance measures, was used for this purpose.

PG4 motif searching, genomic mapping, and analysis

Potential G4 motifs (i.e., G-quadruplex-forming sequences) in the 18 genomes were searched with a customized program written using Perl. We adopted a general pattern: G_n-N_{L1}-G_n-N_{L2}-G_n-N_{L3}-G_n, where G is guanine and N is any nucleotide including G. The number of guanines constituting the stem of G4 DNA (Fig. 1) is given by *n*. *n* was varied from 2 to 5 but restricted to be constant within a particular motif. This does not exclude identification of G4 sequence with variable G-runs as Gs were included in loops, and enables ready detection of the number of tetrads possible in a given sequence (details of the parsing method, includ-

ing comparison with previously published methods, is discussed in the Supplemental material). The number of nucleotides in the three loops L1, L2, and L3 was allowed to vary from 1 to 5, such that the size of loops may vary within a given G4 motif. The program was rerun with cytosine instead of guanine to identify motifs on the − strand and appropriately corrected for orientation before mapping their position in the context of genes. We restricted our program to the above values of stem and loop length after considering the following points. First, single G-tetrads have been observed only in very high (millimolar) guanine solutions and may not be physiologically relevant (Gellert et al. 1962), and a tetrad length exceeding five guanines was not found by us except in only one or two cases. Similar observations have been made before (Todd et al. 2005). We included a stem size of two guanines as various previous reports indicated G4 DNA and RNA with two tetrads as biologically relevant (see Supplemental material) (Wells et al. 1988; Darnell et al. 2001). The loop length was arbitrarily restrained to a maximum of 5 nt for practical reasons. An unrestrained loop length would make searching difficult; moreover, we found that G4 motifs exist as short nucleic acids (a length between 10 and 39 bases was predominant) (Supplemental Table S4), which is also supported by earlier evidence (Hazel et al. 2004). Considering the possible variability in the loops within a motif, we analyzed the loop distribution using a mosaic plot (Friendly 1994) wherein the predominant loop distributions can be readily identified. A single putative quadruplex sequence may present multiple quadruplex topologies with variation in both loop and tetrad compositions. Furthermore, overlapping patterns may be present with more than four G-runs where only one motif is possible at a time (Supplemental Fig. S1). This complicates the analysis for exactly determining the number of possible motifs in a given sequence. We have addressed this by stitching overlapping patterns to present tracts. This tract information has been used for all genomic comparative analysis. However, for analysis of tetrad size and loop combinations, all possible PG4 motifs were considered. This was particularly done to check the prevalence of any structural type or tetrad/loop combination of G4 DNA on a genome-wide scale.

We divided each genome into three regions for mapping of the PG4 motifs: (1) intragenic; (2) putative regulatory (up to 200 bases) upstream of the gene's start codon; and (3) "rest-of-intergenic," comprising all other non-coding intergenic regions (including the downstream intergenic region separating convergently oriented genes). Region 2 comprises the actual intergenic distance when two genes are separated by <200 bases. This partitioning was used for all analysis except where mentioned. The relative abundance of PG4 motifs in different genomic regions was statistically compared, and the significance levels were estimated using the nonparametric Signed Wilcoxon Ranks Test (Wilcoxon 1945). As a control for the significance of PG4 motifs vis-à-vis their distinct genomic distribution, we searched for sequence patterns with multiple G-runs, which were restricted such that they would be unable to adopt a G4 motif (see Supplemental material). The programs written for genome-wide searching, mapping, and analyzing PG4 motifs are available upon request.

In prokaryotes, it is difficult to predict regulatory regions. For a gene within an operon, its promoter may be several genes upstream. In certain cases, a gene within an operon may have its own promoter also. It is difficult to predict operons, and moreover, the first gene in an operon in less well-studied organisms (McGuire et al. 2000). On the other hand, we noticed that a majority of operons in *E. coli* consist of only two to three genes (Salgado et al. 2004). For comparative function analyses across all organisms, we considered the near upstream region of all ORFs as the putative regulatory region. The upstream distance was taken to be −200 bp as the majority of binding sites for DNA-binding proteins in bacteria are found within the first 200 bases upstream of the start codon. Even in cases when there is a site further upstream, one finds another site close to the promoter (Gralla and Collado-Vides 1996). Similarly, very few regulatory sites are present downstream of the start codon; hence, we analyzed only the non-coding regions proximal to the start codon for PG4 motifs, as motifs in this region are expected to be most relevant in gene regulation.

Function classification of genes with PG4 motifs in regulatory region in 18 organisms

We considered 61,974 ORFs in 18 organisms classified in 23 functional classes as defined by the COGS database (Tatusov et al. 1997) for this analysis. All ORFs with one or more PG4 motif(s) within 200 bp of the start codon in each function class were identified. These data were analyzed in two ways after excluding genes that were either not present or not well defined (i.e., general function prediction and function unknown) in COGS. First, we found the percentage of positive hits (genes with PG4 motif in −200 bp) in each class. Second, we found out the function classes with a higher number of positive hits. As a control, we also analyzed the intragenic PG4 motif density of all the 61,974 ORFs. First, the intragenic PG4 motifs were identified, and then the number of bases constituting a motif was counted in each case. This was expressed as "per kb of gene sequence" (motif frequency). The value of motif frequency across all genes in each function class was evaluated for comparative analysis.

Identification of orthologs with conserved PG4 motifs in regulatory region

For finding orthologous genes with conserved PG4 motifs in regulatory regions within the 18 bacterial organisms studied by us, we used the COGS database. The *E. coli* genes with PG4 motifs within the −200-bp region were considered as the reference in this case and used to search for orthologs in the 17 other organisms. Identification of one or more PG4 motif(s) within the −200 region of the ortholog was considered a positive hit, and these genes were classified into the phylogenetic groups defined by STRING server (von Mering et al. 2003). Identification of orthologs with conserved PG4 motif(s) in the regulatory region in distantly related organisms was considered to be significant.

Mapping of PG4 motifs to the transcription regulatory network in *E. coli*

We mapped the regulatory regions with PG4 motifs in *E. coli* in the context of its regulatory network, that is, characterized and predicted operons and promoters. Characterized promoters for 16 genes were obtained from the PromEC database (http://margalit.huji.ac.il/promec/) (Hershberg et al. 2001), and 178 predicted operons and 151 predicted promoters from RegulonDB (http://www.cifn.unam.mx/Computational_Genomics/regulondb/) were used (Salgado et al. 2004).

Mapping of transcription-factor-binding sites to PG4 motifs in the −200-bp region of genes in *E. coli*

We searched for transcription-factor-binding sites (TFBS) associated with all PG4 motifs found in the regulatory region of *E. coli* genes. TFBS (DNA sequence matrices) for 55 *E. coli* DNA-binding proteins from http://arep.med.harvard.edu/ecoli_matrices/ (Robison et al. 1998) with high confidence scores, which comprise

both functionally characterized as well as predicted sites, were used. Sites from the library of matrices were validated against the gene list file for *E. coli* from NCBI (http://www.ncbi.nlm.nih.gov/genomes/) before mapping to flanking regions (100 bases) of PG4 motifs, which are present within 200 bp upstream of the start codon. A window size of 100 bases was used as the flanking region based on previous studies indicating that formation of non-B DNA motifs may have a regulatory effect at a distant promoter site (Hatfield and Benham 2002); however, this distance can be >100 bases. Similarly, we mapped TFBS on 445 putative regulatory regions (comprising 200 bp upstream of start codons) that did not harbor any PG4 motifs, as a control set. All matrices were mapped as reported in the database, without introducing any change for overlapping target sites.

Circular dichroism

Circular dichroism (CD) measurements were performed on a Jasco Spectropolarimeter (model J 715) as described previously (Mathur et al. 2004). See also Supplemental material.

Acknowledgments

We are grateful to Samir K. Brahmachari for support and encouragement. S.C. thanks Partha P. Majumder of the Indian Statistical Institute, Kolkata for helpful discussion; Munia Ganguli, IGIB, for critical reading of the manuscript; and all members of the Chowdhury Lab for useful discussions. K.H. acknowledges a research fellowship (JRF) from CSIR. This work was supported by grants from the CSIR task force project CMM 0017. We also thank the referees for assisting us in enriching the manuscript content.

References

Arimondo, P.B., Riou, J.F., Mergny, J.L., Tazi, J., Sun, J.S., Garestier, T., and Helene, C. 2000. Interaction of human DNA topoisomerase I with G-quartet structures. *Nucleic Acids Res.* **28:** 4832–4838.

Bachrati, C.Z. and Hickson, I.D. 2003. RecQ helicases: Suppressors of tumorigenesis and premature aging. *Biochem. J.* **374:** 577–606.

Bacolla, A. and Wells, R.D. 2004. Non-B DNA conformations, genomic rearrangements, and human disease. *J. Biol. Chem.* **279:** 47411–47414.

Balagurumoorthy, P. and Brahmachari, S.K. 1994. Structure and stability of human telomeric sequence. *J. Biol. Chem.* **269:** 21858–21869.

Balke, V.L. and Gralla, J.D. 1987. Changes in the linking number of supercoiled DNA accompany growth transitions in *Escherichia coli*. *J. Bacteriol.* **169:** 4499–4506.

Beer, M.A. and Tavazoie, S. 2004. Predicting gene expression from sequence. *Cell* **117:** 185–198.

Bentley, S.D., Chater, K.F., Cerdeno-Tarraga, A.M., Challis, G.L., Thomson, N.R., James, K.D., Harris, D.E., Quail, M.A., Kieser, H., Harper, D., et al. 2002. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* **417:** 141–147.

Bordes, P., Conter, A., Morales, V., Bouvier, J., Kolb, A., and Gutierrez, C. 2003. DNA supercoiling contributes to disconnect σ^S accumulation from σ^S-dependent transcription in *Escherichia coli*. *Mol. Microbiol.* **48:** 561–571.

Cheung, I., Schertzer, M., Rose, A., and Lansdorp, P.M. 2002. Disruption of dog-1 in *Caenorhabditis elegans* triggers deletions upstream of guanine-rich DNA. *Nat. Genet.* **31:** 405–409.

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.

Christiansen, J., Kofod, M., and Nielsen, F.C. 1994. A guanosine quadruplex and two stable hairpins flank a major cleavage site in insulin-like growth factor II mRNA. *Nucleic Acids Res.* **22:** 5709–5716.

Cummings, C.J. and Zoghbi, H.Y. 2000. Trinucleotide repeats: Mechanisms and pathophysiology. *Annu. Rev. Genomics Hum. Genet.* **1:** 281–328.

Dapic, V., Abdomerovic, V., Marrington, R., Peberdy, J., Rodger, A., Trent, J.O., and Bates, P.J. 2003. Biophysical and biological properties of quadruplex oligodeoxyribonucleotides. *Nucleic Acids Res.* **31:** 2097–2107.

Darnell, J.C., Jensen, K.B., Jin, P., Brown, V., Warren, S.T., and Darnell, R.B. 2001. Fragile X mental retardation protein targets G quartet mRNAs important for neuronal function. *Cell* **107:** 489–499.

Dunnick, W., Hertz, G.Z., Scappino, L., and Gritzmacher, C. 1993. DNA sequences at immunoglobulin switch region recombination sites. *Nucleic Acids Res.* **21:** 365–372.

Duquette, M.L., Handa, P., Vincent, J.A., Taylor, A.F., and Maizels, N. 2004. Intracellular transcription of G-rich DNAs induces formation of G-loops, novel structures containing G4 DNA. *Genes & Dev.* **18:** 1618–1629.

Etzioni, S., Yafe, A., Khateb, S., Weisman-Shomer, P., Bengal, E., and Fry, M. 2005. Homodimeric MyoD preferentially binds tetraplex structures of regulatory sequences of muscle-specific genes. *J. Biol. Chem.* **280:** 26805–26812.

Florquin, K., Saeys, Y., Degroeve, S., Rouze, P., and Vande, P.Y. 2005. Large-scale structural analysis of the core promoter in mammalian and plant genomes. *Nucleic Acids Res.* **33:** 4255–4264.

Friendly, M. 1994. Mosaic displays for multi-way contingency tables. *J. Am. Stat. Assoc.* 190–200.

Gellert, M., Lipsett, M.N., and Davies, D.R. 1962. Helix formation by guanylic acid. *Proc. Natl. Acad. Sci.* **48:** 2013–2018.

Gilbert, D.E. and Feigon, J. 1999. Multistranded DNA structures. *Curr. Opin. Struct. Biol.* **9:** 305–314.

Gralla, J.D. and Collado-Vides, J. 1996. Organization and function of transcription regulatory elements. In Escherichia coli *and* Salmonella*: Molecular and cellular biology* (ed. F.C. Neidhardt), pp. 1232–1245. ASM Press, Washington, DC.

Halder, K. and Chowdhury, S. 2005. Kinetic resolution of bimolecular hybridization versus intramolecular folding in nucleic acids by surface plasmon resonance: Application to G-quadruplex/duplex competition in human *c-myc* promoter. *Nucleic Acids Res.* **33:** 4466–4474.

Halder, K., Mathur, V., Chugh, D., Verma, A., and Chowdhury, S. 2005. Quadruplex–duplex competition in the nuclease hypersensitive element of human c-myc promoter: C to T mutation in C-rich strand enhances duplex association. *Biochem. Biophys. Res. Commun.* **327:** 49–56.

Hanakahi, L.A., Sun, H., and Maizels, N. 1999. High affinity interactions of nucleolin with G-G-paired rDNA. *J. Biol. Chem.* **274:** 15908–15912.

Hatfield, G.W. and Benham, C.J. 2002. DNA topology-mediated control of global gene expression in *Escherichia coli*. *Annu. Rev. Genet.* **36:** 175–203.

Hazel, P., Huppert, J., Balasubramanian, S., and Neidle, S. 2004. Loop-length-dependent folding of G-quadruplexes. *J. Am. Chem. Soc.* **126:** 16405–16415.

Hengge-Aronis, R. 1993. Survival of hunger and stress: The role of rpoS in early stationary phase gene regulation in *E. coli*. *Cell* **72:** 165–168.

Hershberg, R., Bejerano, G., Santos-Zavaleta, A., and Margalit, H. 2001. PromEC: An updated database of *Escherichia coli* mRNA promoters with experimentally identified transcriptional start sites. *Nucleic Acids Res.* **29:** 277.

Horsburgh, B.C., Kollmus, H., Hauser, H., and Coen, D.M. 1996. Translational recoding induced by G-rich mRNA sequences that form unusual structures. *Cell* **86:** 949–959.

Howell, R.M., Woodford, K.J., Weitzmann, M.N., and Usdin, K. 1996. The chicken β-globin gene promoter forms a novel "cinched" tetrahelical structure. *J. Biol. Chem.* **271:** 5208–5214.

Huppert, J.L. and Balasubramanian, S. 2005. Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.* **33:** 2908–2916.

Incles, C.M., Schultes, C.M., Kempski, H., Koehler, H., Kelland, L.R., and Neidle, S. 2004. A G-quadruplex telomere targeting agent produces p16-associated senescence and chromosomal fusions in human prostate cancer cells. *Mol. Cancer Ther.* **3:** 1201–1206.

Ishihama, A. 1999. Modulation of the nucleoid, the transcription apparatus, and the translation machinery in bacteria for stationary phase survival. *Genes Cells* **4:** 135–143.

Jishage, M., Iwata, A., Ueda, S., and Ishihama, A. 1996. Regulation of RNA polymerase σ subunit synthesis in *Escherichia coli*: Intracellular levels of four species of σ subunit under various growth conditions. *J. Bacteriol.* **178:** 5447–5451.

Kanehisa, M. and Goto, S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28:** 27–30.

Kolodner, R.D. 1995. Mismatch repair: Mechanisms and relationship to cancer susceptibility. *Trends Biochem. Sci.* **20:** 397–401.

Kusano, S., Ding, Q., Fujita, N., and Ishihama, A. 1996. Promoter selectivity of *Escherichia coli* RNA polymerase E σ^{70} and E σ^{38}

holoenzymes. Effect of DNA supercoiling. *J. Biol. Chem.* **271:** 1998–2004.
Martinez-Antonio, A. and Collado-Vides, J. 2003. Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr. Opin. Microbiol.* **6:** 482–489.
Mathur, V., Verma, A., Maiti, S., and Chowdhury, S. 2004. Thermodynamics of i-tetraplex formation in the nuclease hypersensitive element of human c-myc promoter. *Biochem. Biophys. Res. Commun.* **320:** 1220–1227.
McGuire, A.M., Hughes, J.D., and Church, G.M. 2000. Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res.* **10:** 744–757.
McMurray, C.T. 1999. DNA secondary structure: A common and causative factor for expansion in human disease. *Proc. Natl. Acad. Sci.* **96:** 1823–1825.
Modrich, P. and Lahue, R. 1996. Mismatch repair in replication fidelity, genetic recombination, and cancer biology. *Annu. Rev. Biochem.* **65:** 101–133.
Neidle, S. and Read, M.A. 2000. G-quadruplexes as therapeutic targets. *Biopolymers* **56:** 195–208.
Opel, M.L. and Hatfield, G.W. 2001. DNA supercoiling-dependent transcriptional coupling between the divergently transcribed promoters of the ilvYC operon of *Escherichia coli* is proportional to promoter strengths and transcript lengths. *Mol. Microbiol.* **39:** 191–198.
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.
Parkinson, G.N., Lee, M.P., and Neidle, S. 2002. Crystal structure of parallel quadruplexes from human telomeric DNA. *Nature* **417:** 876–880.
Pedersen, A.G., Jensen, L.J., Brunak, S., Staerfeldt, H.H., and Ussery, D.W. 2000. A DNA structural atlas for *Escherichia coli*. *J. Mol. Biol.* **299:** 907–930.
Perez-Martin, J. and de Lorenzo, V. 1997. Clues and consequences of DNA bending in transcription. *Annu. Rev. Microbiol.* **51:** 593–628.
Peter, B.J., Arsuaga, J., Breier, A.M., Khodursky, A.B., Brown, P.O., and Cozzarelli, N.R. 2004. Genomic transcriptional response to loss of chromosomal supercoiling in *Escherichia coli*. *Genome Biol.* **5:** R87.
Rezler, E.M., Seenisamy, J., Bashyam, S., Kim, M.Y., White, E., Wilson, W.D., and Hurley, L.H. 2005. Telomestatin and diseleno sapphyrin bind selectively to two different forms of the human telomeric G-quadruplex structure. *J. Am. Chem. Soc.* **127:** 9439–9447.
Rich, A. and Zhang, S. 2003. Timeline: Z-DNA: The long road to biological function. *Nat. Rev. Genet.* **4:** 566–572.
Robison, K., McGuire, A.M., and Church, G.M. 1998. A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. *J. Mol. Biol.* **284:** 241–254.
Salgado, H., Gama-Castro, S., Martinez-Antonio, A., Diaz-Peredo, E., Sanchez-Solano, F., Peralta-Gil, M., Garcia-Alonso, D., Jimenez-Jacinto, V., Santos-Zavaleta, A., Bonavides-Martinez, C., et al. 2004. RegulonDB (version 4.0): Transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acids Res.* **32:** D303–D306.
Schaffitzel, C., Berger, I., Postberg, J., Hanes, J., Lipps, H.J., and Pluckthun, A. 2001. In vitro generated antibodies specific for telomeric guanine-quadruplex DNA react with *Stylonychia lemnae* macronuclei. *Proc. Natl. Acad. Sci.* **98:** 8572–8577.
Schneider, R., Travers, A., Kutateladze, T., and Muskhelishvili, G. 1999. A DNA architectural protein couples cellular physiology and DNA topology in *Escherichia coli*. *Mol. Microbiol.* **34:** 953–964.
Schneider, R., Travers, A., and Muskhelishvili, G. 2000. The expression of the *Escherichia coli* fis gene is strongly dependent on the superhelical density of DNA. *Mol. Microbiol.* **38:** 167–175.
Seenisamy, J., Rezler, E.M., Powell, T.J., Tye, D., Gokhale, V., Joshi, C.S., Siddiqui-Jain, A., and Hurley, L.H. 2004. The dynamic character of the G-quadruplex element in the c-MYC promoter and modification by TMPyP4. *J. Am. Chem. Soc.* **126:** 8702–8709.
Sen, D. and Gilbert, W. 1988. Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. *Nature* **334:** 364–366.
Shen, J.C. and Loeb, L.A. 2000. The Werner syndrome gene: The molecular basis of RecQ helicase-deficiency diseases. *Trends Genet.* **16:** 213–220.
Sheridan, S.D., Benham, C.J., and Hatfield, G.W. 1999. Inhibition of DNA supercoiling-dependent transcriptional activation by a distant B-DNA to Z-DNA transition. *J. Biol. Chem.* **274:** 8169–8174.
Siddiqui-Jain, A., Grand, C.L., Bearss, D.J., and Hurley, L.H. 2002. Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc. Natl. Acad. Sci.* **99:** 11593–11598.
Simonsson, T., Pecinka, P., and Kubista, M. 1998. DNA tetraplex formation in the control region of c-myc. *Nucleic Acids Res.* **26:** 1167–1172.
Sinden, R.R. 1994. *DNA structure and function*. Academic Press, San Diego.
———. 1999. Biological implications of the DNA structures associated with disease-causing triplet repeats. *Am. J. Hum. Genet.* **64:** 346–353.
Singh, J., Mukerji, M., and Mahadevan, S. 1995. Transcriptional activation of the *Escherichia coli* bgl operon: Negative regulation by DNA structural elements near the promoter. *Mol. Microbiol.* **17:** 1085–1092.
Stover, C.K., Pham, X.Q., Erwin, A.L., Mizoguchi, S.D., Warrener, P., Hickey, M.J., Brinkman, F.S., Hufnagle, W.O., Kowalik, D.J., Lagrou, M., et al. 2000. Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. *Nature* **406:** 959–964.
Strand, M., Prolla, T.A., Liskay, R.M., and Petes, T.D. 1993. Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature* **365:** 274–276.
Tatusov, R.L., Koonin, E.V., and Lipman, D.J. 1997. A genomic perspective on protein families. *Science* **278:** 631–637.
Todd, A.K., Johnston, M., and Neidle, S. 2005. Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res.* **33:** 2901–2907.
Travers, A., Schneider, R., and Muskhelishvili, G. 2001. DNA supercoiling and transcription in *Escherichia coli*: The FIS connection. *Biochimie* **83:** 213–217.
von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B. 2003. STRING: A database of predicted functional associations between proteins. *Nucleic Acids Res.* **31:** 258–261.
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.
Wasserman, W.W., Palumbo, M., Thompson, W., Fickett, J.W., and Lawrence, C.E. 2000. Human–mouse genome comparisons to locate regulatory sites. *Nat. Genet.* **26:** 225–228.
Weinstein-Fischer, D., Elgrably-Weiss, M., and Altuvia, S. 2000. *Escherichia coli* response to hydrogen peroxide: A role for DNA supercoiling, topoisomerase I and Fis. *Mol. Microbiol.* **35:** 1413–1420.
Weitzmann, M.N., Woodford, K.J., and Usdin, K. 1997. DNA secondary structures and the evolution of hypervariable tandem arrays. *J. Biol. Chem.* **272:** 9517–9523.
Wells, R.D., Collier, D.A., Hanvey, J.C., Shimizu, M., and Wohlrab, F. 1988. The chemistry and biology of unusual DNA structures adopted by oligopurine.oligopyrimidine sequences. *FASEB J.* **2:** 2939–2949.
Wilcoxon, F. 1945. Individual comparisons by ranking methods. *Biometrics* 80–83.
Woodford, K.J., Howell, R.M., and Usdin, K. 1994. A novel K+-dependent DNA synthesis arrest site in a commonly occurring sequence motif in eukaryotes. *J. Biol. Chem.* **269:** 27029–27035.
Wu, X. and Maizels, N. 2001. Substrate-specific inhibition of RecQ helicase. *Nucleic Acids Res.* **29:** 1765–1771.
Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., and Kellis, M. 2005. Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434:** 338–345.
Zahler, A.M., Williamson, J.R., Cech, T.R., and Prescott, D.M. 1991. Inhibition of telomerase by G-quartet DNA structures. *Nature* **350:** 718–720.

Received August 25, 2005; accepted in revised form March 2, 2006.