QuadBase: genome-wide database of G4 DNA—occurrence and conservation in human, chimpanzee, mouse and rat promoters and 146 microbes

Vinod Kumar Yadav[1], James Kappukalayil Abraham[1], Prithvi Mani[1], Rashi Kulshrestha[1] and Shantanu Chowdhury[1,2,*]

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

Received August 9, 2007; Revised September 15, 2007; Accepted September 17, 2007

ABSTRACT

Emerging evidence indicates the importance of G-quadruplex motifs as drug targets. [Stuart A. Borman, Ascent of quadruplexes—nucleic acid structures become promising drug targets. *Chem. Eng. News*, 2007;85, 12–17], which stems from the fact that these motifs are present in a surprising number of promoters wherein their role in controlling gene expression has been demonstrated for a few. We present a compendium of quadruplex motifs, with particular focus on their occurrence and conservation in promoters—QuadBase. It is composed of two parts (EuQuad and ProQuad). EuQuad gives information on quadruplex motifs present within 10 kb of transcription starts sites in 99 980 human, chimpanzee, rat and mouse genes. ProQuad contains quadruplex information of 146 prokaryotes. Apart from gene-specific searches for quadruplex motifs, QuadBase has a number of other modules. 'Orthologs Analysis' queries for conserved motifs across species based on a selected reference organism; 'Pattern Search' can be used to fetch specific motifs of interest from a selected organism using user-defined criteria for quadruplex motifs, i.e. stem, loop size, etc. 'Pattern Finder' tool can search for motifs in any given sequence. QuadBase is freely available to users from non-profit organization at http://quadbase.igib.res.in/.

INTRODUCTION

Biological relevance of non-B DNA motifs in recombination, replication and particularly, regulation of gene expression has been generally appreciated in recent years (1–4). However, the mechanisms underlying the involvement of non-B DNA motifs in function are not clearly understood. It is increasingly becoming 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 a way where DNA sequence information *per se* has minimal role besides facilitating formation of the structural motifs. In the context of gene regulation, several reports have implicated the role of non-B DNA, both in prokaryotes (5) and eukaryotes (3,6). In order to better understand the role of non-B DNA motifs, with particular emphasis on transcription we used the quadruplex or G4 DNA motif as a case study.

Guanine-rich sequences attain unique four-stranded conformations known as G4 DNA (7–9). G4 DNA stabilized by charge coordination with monovalent cations (especially K^+) between a planar array of four hydrogen-bonded guanines (G-quartets or tetrads) may result from intra- or intermolecular association of four DNA strands in parallel or antiparallel orientation (Figure 1) (10). Chromosomal regions containing guanine-rich sequence capable of forming G4 DNA include immunoglobin heavy chain switch regions (11), G-rich minisatellites (12,13), rDNA (14) and telomeres (15), where it plays a central role in telomere extension and is the focus of targeted anticancer drug development (16–18).

Escherichia coli RecQ can unwind G4 DNA and the family of RecQ helicases is conserved and is essential for genomic stability in organisms from *E. coli* to humans (19–21). On the 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 (5,22,23). In a regulatory context, emerging evidence shows several important gene promoters like β-globin genes (24), retinoblastoma susceptibility genes (25), the

*To whom correspondence should be addressed. Fax: +91 11 2766 7471; Email: shantanuc@igib.res.in insulin gene (26), adenovirus serotype 2 (27), *PDGF* (28), *c-KIT* (29), hypoxia inducible factor 1-alpha (30), *BCL-2* (31) and *c-MYC* (32,33) harbor G4 DNA motifs with possible functional role. Repression of *c-MYC* on stabilization of G4-motif (using the G4-binding ligand TMPyP4) and over-expression of *c-MYC* in case of site-specific mutations that destabilized a G4-motif within the *c-MYC* promoter indicated functional role of G4 DNA in transcription (33). These studies largely implicate G4-motifs in a regulatory role on a case-to-case basis warranting investigation of G4 DNA in a global regulatory role. We recently studied this aspect in both prokaryotes and eukaryotes—a genome-wide analysis of 18 microbes indicated enrichment of G4 DNA motifs in putative promoters wherein detailed analysis in *E. coli* suggested global role of G4 motifs in 'turning-on' transcription during growth phase (34). Recent studies have noted the prevalence of quadruplex motifs within the human genome, particularly within promoters (35–37). A study comprising all human, chimpanzee, mouse and rat genes showed that G4 DNA motifs are enriched and conserved (interestingly, more than 700 human promoters conserve G4 DNA motifs with corresponding mouse and rat promoters) within putative promoters and also share many general characteristics of regulatory elements like, tissue-specific expression and association with core promoter elements. Apart from this, significant change in whole genome expression under the influence of a G4 DNA-binding molecule was observed (Verma *et al.*, unpublished results).

Keeping these aspects in mind, herein we present a collated form of information on G4 DNA sequences in 146 bacterial species and human, chimpanzee, mouse and rat in a relational database that enables motif searches and querying for conserved elements across related species with relative ease. Apart from user-defined inputs on motif type (stem and loop size) and location-of-interest with respect to the gene, additional information on genomic location, type of pattern and gene-specific conservation can be downloaded for further use.

SEARCHING AND MAPPING OF G4 DNA MOTIFS

G4 DNA motifs comprise of four runs of guanines (constituting the stem of G4 motif) interspersed with nucleotide bases, which form three intervening loops (Figure 1). We developed a pattern search algorithm to identify potential G4 DNA sequences wherein four consecutive G-runs were identified, considering three intervening loops. 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 be present as one exclusive motif, as described earlier (34). 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: (i) occurrence of the patterns or tracts and (ii) association of the tracts with different genomic regions. PG4 motifs in 146 prokaryotic genomes and four eukaryotic genomes were searched with a customized program written using Java, which adopted a general pattern: Gn-NL1-Gn-NL2-Gn-NL3-Gn, where G is guanine and N is any nucleotide including G. The number of guanines constituting the stem of G4 DNA is given by *n*. The value of *n* is 3 and constant within a particular motif. The number of nucleotides in the three loops L1, L2 and L3 was allowed to vary from 1 to 7, such that the size of loops may vary within a given G4 DNA motif.

A total of 146 bacterial organisms were used for our analysis after downloading their genomes from the NCBI database (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi). Human, mouse, rat and chimpanzee genomic data was downloaded from http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi. Human build 35, mouse build 34, rat build 3.4 and chimpanzee build 1 were used for analysis.

DATABASE ARCHITECTURE

A relational database was constructed in MySQL for storage and query of data. It includes two key entities: 'Pattern Search' and 'Orthologs Analysis' (Figure 2), which are related on the basis of gene name. This was implemented independently for 146 prokaryotes (ProQuad) and human, mouse, rat and chimpanzee (EuQuad). Pattern Search retrieves information on quadruplex. It searches the database on the basis of selected organism(s) and other query parameters such as pattern type, genomic location, loop and stem size. Orthologs Analysis on the other hand retrieves information on conserved quadruplex motifs within selected genomic regions in organisms of interest. Herein, once genes of interest along with reference organism are selected, a query is searched on organisms wherein one intends to find conservation with parameters that define quadruplex motifs such as pattern type, loop size, etc. In EuQuad, this query is with reference to human only and a graphical representation of conserved G4 DNA on promoters is also presented (Figure 3). Additionally, we have included a tool box that enables one to find PG4 motifs in any given sequence. All outputs are given both as html and downloads.

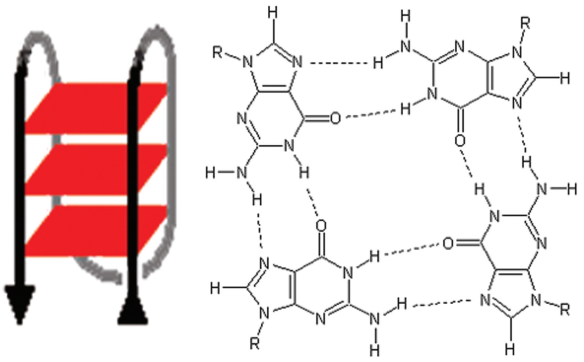


Figure 1. Schematic representation of G4 motif. Red planes show three G-tetrads constituting the stem (left panel) and hydrogen bonding scheme of guanines forming tetrad is shown in right panel.

Pattern search

Pattern Search is an interface to query quadruplex sequences of 146 prokaryotes and 4 eukaryotes (Figure 2). In ProQuad we divided each genome into three regions for mapping of the PG4 motifs: (i) intragenic, (ii) putative regulatory, up to 200 bases (or as defined by user) upstream of the genes start codon and (iii) rest-of-intergenic, comprising all other non-coding intergenic regions (including the downstream intergenic region separating convergently oriented genes). Putative regulatory region comprises the actual intergenic distance when two genes are separated by <200 bases. Query can be performed by selecting organism, pattern type, stem size, loop size and also by providing NCBI Gene ID(s) separated by commas in Gene name field. Each identified pattern can be further used to find all possible loop combinations. This application will enable detailed analysis of variation in loop topology, which is known to result in different structural forms (38).

EuQuad comprises tables for quadruplex information in human, rat, mouse and chimpanzee promoters [10 kb region centered at transcription start site (TSS)]. In Pattern Search, PG4 motifs were searched with a customized algorithm, as described above. We divided each chromosome into two regions for mapping of the PG4 motifs: (i) upstream (ii) downstream with respect to the TSS.

Orthologs analysis

In ProQuad the main objective is to find the PG4 motifs that are conserved between orthologous genes across organisms using the COG (http://www.ncbi.nlm.nih.gov/COG/) database for basic information. One organism is selected as the reference and others (to a maximum of 20 organisms) can be queried in parallel. The generated output listed gene-wise along with details of pattern size, location, etc. are given both as an html file and an excel worksheet for downloading.

In EuQuad PG4 motif conservation is searched in rat, mouse and chimpanzee with human as a reference organism. This search is restricted to putative regulatory regions (4 kb region centered at TSS). For analysis of conservation a window size centered on the quadruplex position in human is given for searching in other genomes—e.g. a window size of 50 bases will search for conserved occurrence within ± 50 bases of the position in human (with respect to TSS). This window size can be user-defined. A viewer interface shows the conserved motifs in selected organisms (Figure 3).

Our program detects PG4 motifs that are conserved across organisms. This conservation is essentially at the structure level, whereby they are expected to result in G4 motifs. In other words, all sequences that conform to the G4 pattern defined by user will be identified, the primary sequences may not be identical. Additionally, variation in loop size, length and sequence leading to different G4 motif topologies and resulting thermodynamic stability issues are discussed in several recent studies (38).

Pattern finder

This is a tool that allows one to find quadruplex motif in a given sequence of interest. Search can be designed based

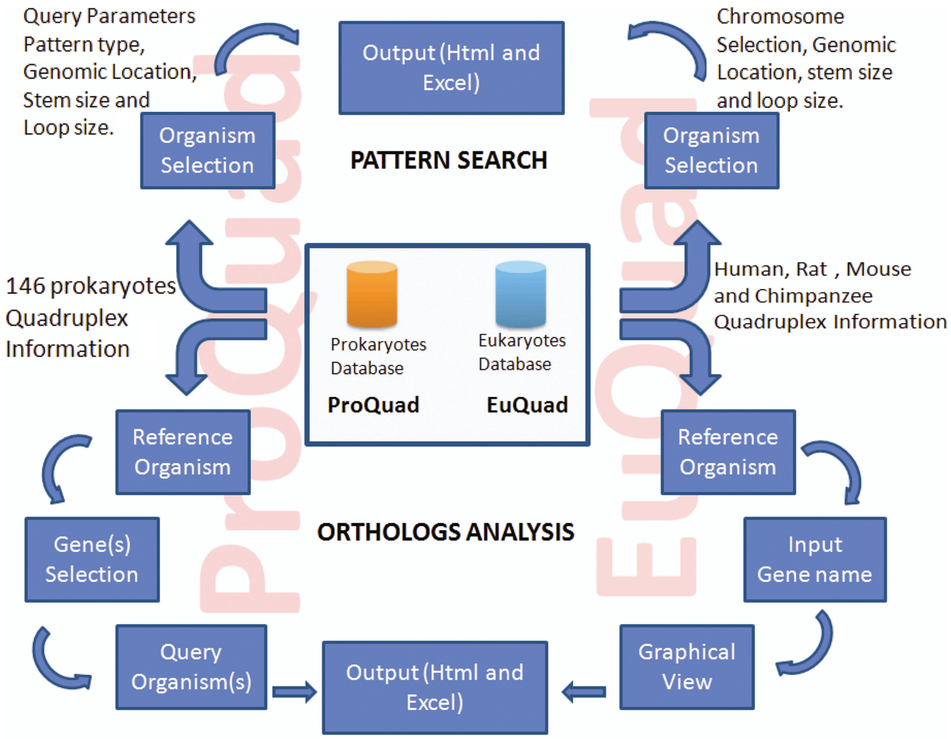


Figure 2. Database architecture for EuQuad and ProQuad. Relational database containing PG4 motif information of human, rat, mouse, chimpanzee and 146 prokaryotes. In Pattern Search, organism is selected with other query parameters such as pattern type, stem size, genomic location and loop size. Orthologs Analysis retrieves information based on query fired for selected genes of selected reference organism. All output is in html as well as in downloadable format.

on various parameters—pattern type G or C in stem (i.e. quadruplex of i-motif sequences, respectively can be retrieved), stem and loop size. Results can be downloaded in spreadsheet format. Nucleotide sequences can be uploaded or data can be posted in a text field provided in the page. Other tools available for finding G4 DNA motifs are QuadFinder and QGRS Mapper. QuadFinder is a tool where quadruplex motifs are searched in given sequence, similarly we have pattern finder. Here we also incorporate download feature for output in our tool. QGRS Mapper is a tool to find Quadruplex forming G-rich sequences in human genes. In our database we cover human, rat, mouse and chimpanzee, we also provide feature to look conservation of human G4 DNA in mouse, rat and chimpanzee orthologs with user-defined window.

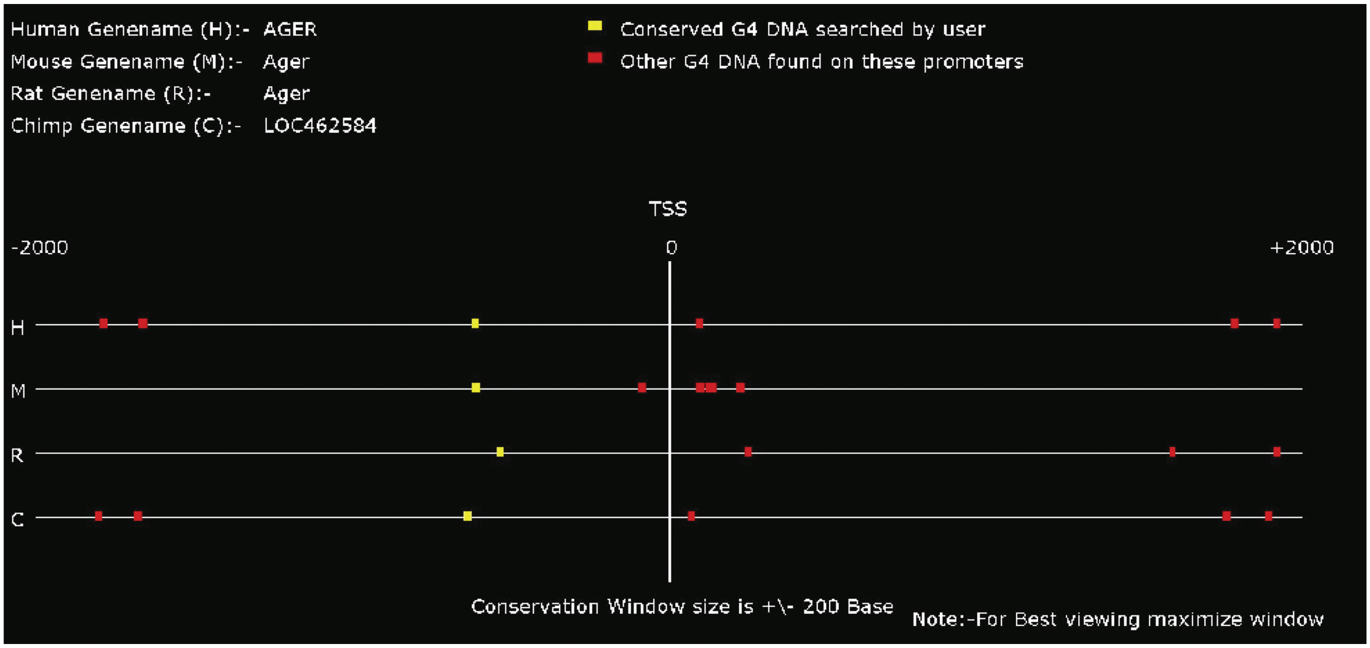
CONCLUSION

In the current state we have focused on making an initial compendium of the quadruplex data, which primarily caters to transcription and kept the focus, from a user's point of view, on conservation analysis. The presentation is G4 DNA-centric and does not incorporate other forms of information like transcription factor-binding site or functional classification of genes. At this state QuadBase presents two separate databases for prokaryotes and eukaryotes and navigation across these two databases is not possible. The G4 DNA pattern executed by our algorithm (see above) is a generic form of the structure and is expected to represent a bulk of G4 DNA motifs found in the genome. A similar pattern has been used by almost all recent studies in order to find motifs genome-wide (37,39). However, other intramolecular forms of G4 DNA reported recently [e.g. the 3 + 1 forms (40,41)] will not be detected by our program.

A previously published database, GRSDB contains quadruplex forming G-rich sequences present in alternatively processed mammalian pre-mRNA sequences (42). GRSDB has 1310 human and mouse genes, of which 1188 are alternatively processed. At least two quadruplex finder application softwares are available (43,44), which allow one to find G4 DNA motifs in given sequences. QuadBase gives all quadruplex motifs present within 10 kb of TSS of 99 980 promoter in human, chimpanzee, mouse and rat and analyses for conserved patterns in 'orthologous genes'. Genome-wide identification of quadruplex motifs in 146 prokaryotes is possible, where one can also analyse conservation of motifs across multiple species. To the best of our knowledge, QuadBase is the first database of G4 DNA motifs in prokaryotes and human and related species, which particularly focuses on conserved motifs that could be important in gene expression. The primary features of EuQuad can be used to quickly query for user-defined patterns on a particular chromosome and ask whether it is conserved in other related organisms. The output in case of multiple such occurrences is a list showing specific type of the pattern, its location in human and other species along with a graphic viewer showing the conserved motifs. Alternatively, one could search one or more genes in a similar query. ProQuad, essentially allows similar applications across 146 microbes in a genome-wide context.

ACKNOWLEDGEMENT

Authors acknowledge funding from CSIR—Task Force Project—CMM 0017 for the entire project. Authors also



Human Genename (H):- AGER
Mouse Genename (M):- Ager
Rat Genename (R):- Ager
Chimp Genename (C):- LOC462584

🟨 Conserved G4 DNA searched by user
🟥 Other G4 DNA found on these promoters

Conservation Window size is +\- 200 Base

Note:-For Best viewing maximize window

Figure 3. Graphical representation of conserved PG4 motif within promoters. Searched motifs that are conserved in rat, mouse and chimpanzee with human as reference organism are shown. Each horizontal line represents promoter sequence (4 kb centered at TSS). Color boxes represent total PG4 motifs in promoter sequence and width represents relative length of the motif. Yellow boxes represent conserved G4 motifs with respect to the one selected.

thank the IGIB systems administration support group for technical help and all other members of the Chowdhury laboratory for helpful discussions. The Open Access publication charges for this article were waived by the Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Sinden,R.R. (1994) *DNA: Structure and Function*. Academic Press, San Diego, CA.
2. Perez-Martin,J. and de Lorenzo,V. (1997) Clues and consequences of DNA bending in transcription. *Annu. Rev. Microbiol.*, **51**, 593–628.
3. Bacolla,A. and Wells,R.D. (2004) Non-B DNA conformations, genomic rearrangements, and human disease. *J. Biol. Chem.*, **279**, 47411–47414.
4. 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.
5. 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.
6. Rich,A. and Zhang,S. (2003) Timeline: Z-DNA: the long road to biological function. *Nat. Rev. Genet.*, **4**, 566–572.
7. 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.
8. Gellert,M., Lipsett,M.N. and Davies,D.R. (1962) Helix formation by guanylic acid. *Proc. Natl Acad. Sci. USA*, **48**, 2013–2018.
9. Balagurumoorthy,P. and Brahmachari,S.K. (1994) Structure and stability of human telomeric sequence. *J. Biol. Chem.*, **269**, 21858–21869.
10. Gilbert,D.E. and Feigon,J. (1999) Multistranded DNA structures. *Curr. Opin. Struct. Biol.*, **9**, 305–314.
11. 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.
12. 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.
13. Jeffreys,A.J., Royle,N.J., Wilson,V. and Wong,Z. (1988) Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. *Nature*, **332**, 278–281.
14. 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.
15. Parkinson,G.N., Lee,M.P. and Neidle,S. (2002) Crystal structure of parallel quadruplexes from human telomeric DNA. *Nature*, **417**, 876–880.
16. 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.
17. Neidle,S. and Read,M.A. (2000) G-quadruplexes as therapeutic targets. *Biopolymers*, **56**, 195–208.
18. 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.
19. Wu,X. and Maizels,N. (2001) Substrate-specific inhibition of RecQ helicase. *Nucleic Acids Res.*, **29**, 1765–1771.
20. 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.
21. Bachrati,C.Z. and Hickson,I.D. (2003) RecQ helicases: suppressors of tumorigenesis and premature aging. *Biochem. J.*, **374**, 577–606.
22. 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.
23. 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.
24. Howell,R.M., Woodford,K.J., Weitzmann,M.N. and Usdin,K. (1996) The chicken beta-globin gene promoter forms a novel ‘‘cinched’’ tetrahelical structure. *J. Biol. Chem.*, **271**, 5208–5214.
25. Murchie,A.I. and Lilley,D.M. (1992) Retinoblastoma susceptibility genes contain 5′ sequences with a high propensity to form guanine-tetrad structures. *Nucleic Acids Res.*, **20**, 49–53.
26. Catasti,P., Chen,X., Moyzis,R.K., Bradbury,E.M. and Gupta,G. (1996) Structure-function correlations of the insulin-linked polymorphic region. *J. Mol. Biol.*, **264**, 534–545.
27. Kilpatrick,M.W., Torri,A., Kang,D.S., Engler,J.A. and Wells,R.D. (1986) Unusual DNA structures in the adenovirus genome. *J. Biol. Chem.*, **261**, 11350–11354.
28. Ma,D., Xing,Z., Liu,B., Pedigo,N.G., Zimmer,S.G., Bai,Z., Postel,E.H. and Kaetzel,D.M. (2002) NM23-H1 and NM23-H2 repress transcriptional activities of nuclease-hypersensitive elements in the platelet-derived growth factor-A promoter. *J. Biol. Chem.*, **277**, 1560–1567.
29. Rankin,S., Reszka,A.P., Huppert,J., Zloh,M., Parkinson,G.N., Todd,A.K., Ladame,S., Balasubramanian,S. and Neidle,S. (2005) Putative DNA quadruplex formation within the human c-kit oncogene. *J. Am. Chem. Soc.*, **127**, 10584–10589.
30. De,A.R., Wood,S., Sun,D., Hurley,L.H. and Ebbinghaus,S.W. (2005) Evidence for the presence of a guanine quadruplex forming region within a polypurine tract of the hypoxia inducible factor 1alpha promoter. *Biochemistry*, **44**, 16341–16350.
31. Dexheimer,T.S., Sun,D. and Hurley,L.H. (2006) Deconvoluting the structural and drug-recognition complexity of the G-quadruplex-forming region upstream of the bcl-2 P1 promoter. *J. Am. Chem. Soc.*, **128**, 5404–5415.
32. Simonsson,T., Pecinka,P. and Kubista,M. (1998) DNA tetraplex formation in the control region of c-myc. *Nucleic Acids Res.*, **26**, 1167–1172.
33. 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. USA*, **99**, 11593–11598.
34. Rawal,P., Kummarasetti,V.B., Ravindran,J., Kumar,N., Halder,K., Sharma,R., Mukerji,M., Das,S.K. and Chowdhury,S. (2006) Genome-wide prediction of G4 DNA as regulatory motifs: role in *Escherichia coli* global regulation. *Genome Res.*, **16**, 644–655.
35. Todd,A.K., Johnston,M. and Neidle,S. (2005) Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res.*, **33**, 2901–2907.
36. Huppert,J.L. and Balasubramanian,S. (2005) Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.*, **33**, 2908–2916.
37. Huppert,J.L. and Balasubramanian,S. (2007) G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res.*, **35**, 406–413.
38. Luu,K.N., Phan,A.T., Kuryavyi,V., Lacroix,L. and Patel,D.J. (2006) Structure of the human telomere in K+ solution: an intramolecular (3 + 1) G-quadruplex scaffold. *J. Am. Chem. Soc.*, **128**, 9963–9970.
39. Burge,S., Parkinson,G.N., Hazel,P., Todd,A.K. and Neidle,S. (2006) Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res.*, **34**, 5402–5415.
40. Dai,J., Punchihewa,C., Ambrus,A., Chen,D., Jones,R.A. and Yang,D. (2007) Structure of the intramolecular human telomeric G-quadruplex in potassium solution: a novel adenine triple formation. *Nucleic Acids Res.*, **35**, 2440–2450.
41. Scaria,V., Hariharan,M., Arora,A. and Maiti,S. (2006) Quadfinder: server for identification and analysis of quadruplex-forming motifs in nucleotide sequences. *Nucleic Acids Res.*, **34**, W683–W685.
42. Kostadinov,R., Malhotra,N., Viotti,M., Shine,R., D'Antonio,L. and Bagga,P. (2006) GRSDB: a database of quadruplex forming G-rich sequences in alternatively processed mammalian pre-mRNA sequences. *Nucleic Acids Res.*, **34**, D119–D124.
43. Zhao,Y., Du,Z. and Li,N. (2007) Extensive selection for the enrichment of G4 DNA motifs in transcriptional regulatory regions of warm blooded animals. *FEBS Lett.*, **581**, 1951–1956.
44. Kikin,O., D'Antonio,L. and Bagga,P.S. (2006) QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res.*, **34**, W676–W682.