2700–2704 *Nucleic Acids Research, 2008, Vol. 36, No. 8* *Published online 19 March 2008*
doi:10.1093/nar/gkn078

The relationship of potential G-quadruplex sequences in *cis*-upstream regions of the human genome to SP1-binding elements

Alan K. Todd and Stephen Neidle*

CRUK Biomolecular Structure Group, The School of Pharmacy, University of London, 29-39 Brunswick Square, London WC1N 1AX, UK

Received January 24, 2008; Revised February 7, 2008; Accepted February 8, 2008

ABSTRACT

We have carried out a survey of potential quadruplex structure sequences (PQSS), which occur in the immediate upstream region (500 bp) of human genes. By examining the number and distribution of these we have established that there is a clear link between them and the occurrence of the SP1-binding element 'GGGCGG', such that a large number of upstream PQSS incorporate the SP1-binding element.

INTRODUCTION

Certain guanine-rich DNA sequences have the ability to form stable secondary structures, G-quadruplexes, comprising G-tetrad motifs that involve four guanines arranged in a planar array interacting via Hoogsteen hydrogen bonds (1,2). These G-tetrads are stable when a number are stacked on top of one another. There have been numerous topologies observed for G-quadruplexes (3) and a number of studies have mapped out potential quadruplex forming sequences in genomic DNAs (4–11). The search criteria for most of these surveys have been sequences, which contain four or more runs of G-tracts occurring close together on the same strand. Most of the observed topologies follow this pattern (12,13). These potential quadruplex structure sequences (PQSS) have been found to occur with elevated frequency in regions directly upstream of the transcription start site of genes in species as diverse as *Escherichia coli* (8) and humans (11). There are also a number of specific promoter regions for which there is biophysical and structural evidence for the formation of PQSSs (for example, 11–18), at least *in vitro*.

The zinc finger protein SP1 acts as a transcription factor, which has been shown to bind to the upstream element sequence 'GGGCGG' (19–21). This element has been found in many different promoters, often with a copy number >1 (19,20) and often within the first 100 bp upstream of the transcription start site. The consensus sequence has been shown to be 'GGGGCGGGGC' (22,23). The fact that it is guanine-rich, with consecutive guanines, gives it the ability to participate in PQSSs, at least in principle. We show here that many of the PQSSs found in the regions directly upstream of the transcription start site actually contain the SP1 consensus sequence and that there is a correlation between genes, which have the SP1 and PQSS in upstream regions.

Another common upstream promoter element 'CCAAT' occurs in a similarly large number of promoters but does not contribute as much guanine-richness as the SP1-binding element and can be employed as a useful comparative group.

METHODS

The mySQL tables for the Ensembl human core database v45 (24) were downloaded from the Ensembl web site (www.ensembl.org) and imported onto a local computer. The human genome was searched for PQSSs in the same way as described earlier (5) and the PQSS data was uploaded into mySQL tables. The search criterion was four runs of guanine $G_m X_n G_m X_o G_m X_p G_m$ where m is between 3 and 5, and X are any combination of bases where n, o and p are between 1 and 7.

Using Perl scripts and the Ensembl Perl API (25), a list of all genes with Ensembl status 'known' was compiled and the 500 bp upstream flanking sequences were extracted. These were searched for the sequences 'GGGCGG' and 'CCAAT' as well as their complementary sequences and the genes were grouped into the following categories:

(i) Genes with 'GGGCGG' in the region 500 bps upstream of their transcription start site (500USR).
(ii) Genes with 'CCAAT' in their 500USR.

*To whom correspondence should be addressed. Tel: 0044 207 753 5969; Fax: 0044 207 753 5970; Email: stephen.neidle@pharmacy.ac.uk

(iii) Genes with both in their 500USR.
(iv) Genes with 'GGGCGG' but not 'CCAAT' in their 500USR.
(v) Genes with 'CCAAT' but not 'GGGCGG' in their 500USR.
(vi) Genes with neither sequence in their 500USR.

The MySQL table of PQSSs was used to count the PQSSs in the upstream regions of the genes in each category and their distances (−1 bp to −500 bps) from the transcription start sites were noted. In order to represent these graphically the positions of the quadruplex sequences were put into bins of 10 base pairs. When looking at the number of PQSSs we counted each contiguous region of potential quadruplex region as a single sequence element i.e. if there were overlaps between more than one PQSS, then these were counted as one sequence element. For example the sequence: **GGG**AG GC**GGG**CGGT**GGGGGGGG**TG**GGG**GTG**GGG**, which is in the upstream region of the Ensembl gene ENSG00000204219 (HGNC symbol TCEA3), was counted as one sequence element. Even though it contains up to six runs of guanines, we assume that only one quadruplex structure at a time can be formed in this region.

The number of cytosines and guanines in all known genes was also derived for each of the bins. The cytosine and guanine data were normalized as was the total quadruplex distribution so that a comparison of the relative distributions could be made, since the absolute number of cytosines and guanines was much higher. Our database of PQSSs was also searched for PQSSs, which incorporated the SP1 consensus sequence and the distributions of PQSSs containing the SP1 consensus sequence and those not containing the SP1 consensus sequence could then be determined.

RESULTS AND DISCUSSION

Figure 1 shows the number of PQSS elements per gene for a number of different grouping of genes. Bars A and B show that PQSSs in upstream regions of genes that contain SP1 elements are much more common than PQSSs in upstream regions of genes without SP1 elements. In bars C and D we see that there is no such relationship in genes whose upstream regions contain CCAAT elements. If anything the reverse is true and genes without CCAAT elements contain more potential PQSSs. This trend is repeated in bars E, F, G and H where E and G contain SP1 elements and contain many more PQSSs per gene than bars F and H. The trend of fewer PQSSs in the upstream regions with CCAAT is also apparent here. In bar I the number of PQSS elements per gene in the upstream region of all genes is much lower than those which contain SP1 sequence elements and higher than those which do not.

Figure 2 shows the distribution of the promoter-binding elements SP1 (GGGCGG) and CCAAT with distance from the transcription start site. The shape of this graph shows that the frequency of sequence elements rises steeply, reaching a peak in the −50 to −41 range for SP1 and in the −70 to −61 range for CCAAT before falling off gradually.

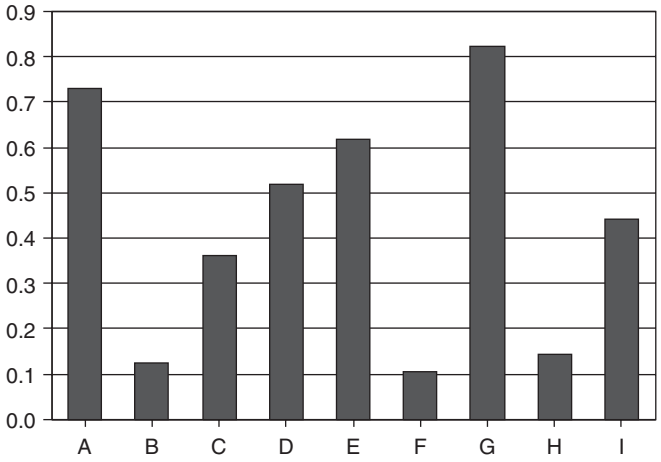

Figure 1. PQSS sequence elements per gene for potential quadruplex sequences in upstream regions (**A**) containing SP1 elements, (**B**) not containing SP1 elements, (**C**) containing CCAAT elements, (**D**) without CCAAT elements, (**E**) containing CCAAT and SP1 elements, (**F**) containing CCAAT and no SP1 elements, (**G**) containing SP1 and no CCAAT elements, (**H**) containing neither SP1 nor CCAAT elements, (**I**) of all known Ensembl genes.

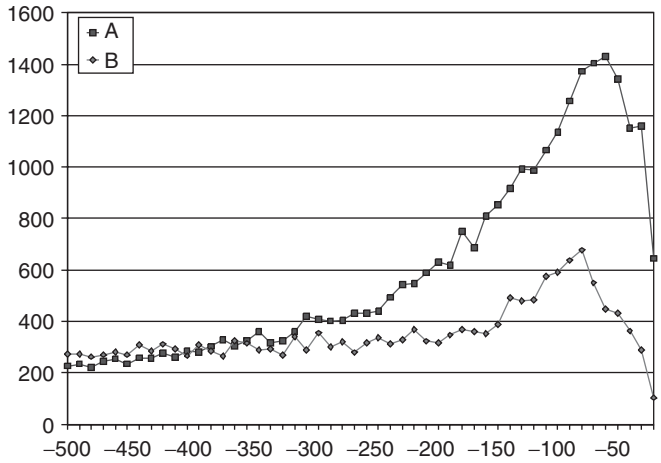

Figure 2. (**A**) SP1 sequence elements, which occur in upstream regions, (**B**) CCAAT sequence elements, which occur in upstream regions.

The distribution of PQSSs in the 500 bp upstream region of transcription start sites is very similar to that of the regulatory motifs in Figure 2. Figure 3A shows the distribution for PQSSs in the upstream region of all Ensembl genes with status 'known'; the maximum peak is in the same region as the peak for the distribution of SP1-binding elements, −50 to −41 bases. A search for the SP1-binding site motif revealed that of the 22 633 known genes, just over half (52.5%) contained the motif in their upstream region (Table 1). However in Figure 3B we can see that this set of genes accounts for the vast majority of PQSSs in upstream regions (86.6%). Not only are the absolute number of PQSSs different, but the distribution of PQSSs which are in the upstream regions of genes with

and without the SP1 consensus sequence differ markedly, as seen in Figure 3B and C, respectively. The PQSSs in non-SP1 upstream regions have a much flatter distribution than that of the SP1 motif genes.

Since the SP1 consensus sequence is guanine-rich and can be incorporated into PQSSs, we examined the number of PQSSs, which contained the SP1 consensus sequence. These account for just under half the total PQSSs (47.2%). The distribution of PQSSs incorporating the SP1 sequence (Figure 3D) and PQSSs without the SP1 consensus sequence (Figure 3E) is very different. Figure 3D shows a distribution similar to the SP1 motif while that in Figure 3E is much flatter.

For a random sequence the probability of finding a PQSS is related to its guanine content so we looked at the guanine density within upstream regions. In Figure 4 we have the normalized distributions of PQSSs (A), guanine bases (B) and cytosine bases (C). Both guanine and cytosine do indeed get more frequent closer to the transcription start site although it is hard to say whether this is related to PQSS distribution.

Figure 5 shows the distributions of PQSSs within genes that contain the regulatory element CCAAT in their upstream region (Figure 5B) and PQSSs within genes, which do not contain CCAAT in their upstream region (Figure 5C). The distribution of PQSSs in CCAAT,

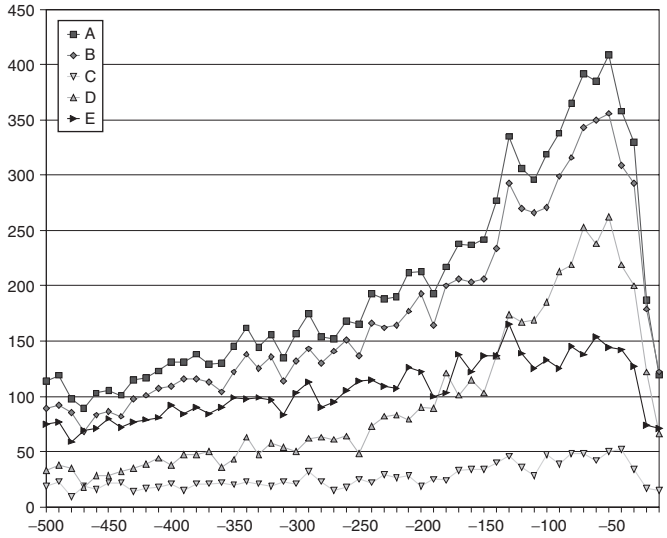

Figure 3. Potential quadruplex sequences (**A**) occurring in upstream regions, (**B**) in upstream regions which contain SP1 sequence elements (**C**) in upstream regions which do not contain SP1 sequence elements, (**D**) which incorporate SP1 sequence elements and are within upstream regions, (**E**) which do not incorporate SP1 sequence elements and are within upstream regions.

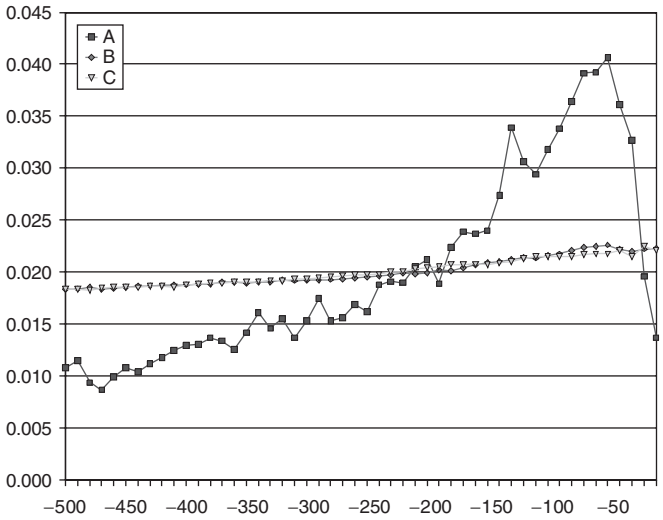

Figure 4. Normalized distributions in the 500 bp upstream regions of Ensembl genes with status 'known' of (**A**) PQSSs, (**B**) guanines, (**C**) cytosines.

Table 1. Summary of quadruplex occurrences in upstream regions of the human genome

Quadruplex occurences	Number of genes	Number of sequence elements
SP1 sequence elements which occur in upstream regions	11 872	29 991
PQSSs in upstream regions which contain SP1 sequence elements	5415	8660
Genes without SP1 consensus sequence in their upstream regions	10 761	
PQSSs in upstream regions which do not contain SP1 sequence elements	1096	1335
PQSSs which incorporate SP1 sequence elements and are within upstream regions	3596	4721
PQSSs which do not incorporate SP1 sequence elements and are within upstream regions	4150	5274
CCAAT sequence elements which occur in upstream regions	10 963	17 574
PQSSs in upstream regions containing CCAAT sequence elements	2718	3948
Genes whose upstream regions contain no CCAAT sequence elements	11 670	
PQSSs in upstream regions containing no CCAAT sequence elements	3794	6047
Genes whose upstream regions contain CCAAT and SP1	5451	
PQSSs in upstream regions containing SP1 and CCAAT sequence elements	2236	3368
Genes whose upstream regions contain CCAAT and no SP1	5512	
PQSSs in upstream regions containing CCAAT and no SP1 sequence elements	482	580
Genes whose upstream regions contain SP1 and no CCAAT	6421	
PQSSs in upstream regions containing SP1 and no CCAAT elements	3180	5292
Genes whose upstream regions contain neither SP1 nor CCAAT	5249	
PQSSs in upstream regions containing neither SP1 nor CCAAT elements	614	755
All genes	22 633	
PQSSs in all upstream regions	6512	9995

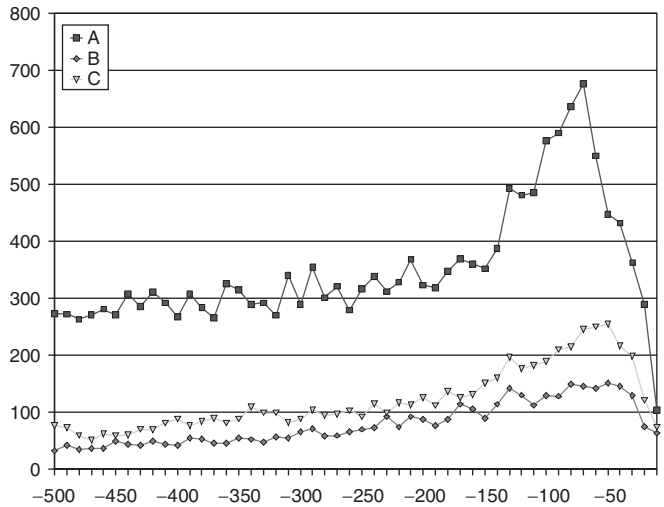

Figure 5. (**A**) CCAAT sequence elements occurring in upstream regions. (**B**) PQSSs in upstream regions containing CCAAT sequence elements. (**C**) PQSSs in upstream regions containing no CCAAT sequence elements.

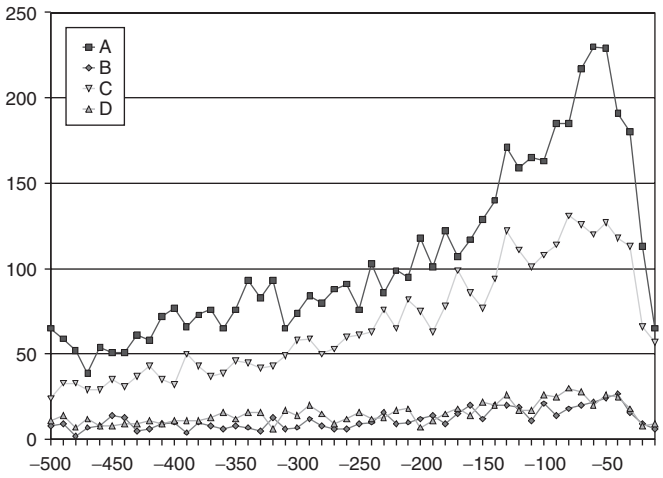

Figure 6. (**A**) PQSSs in upstream regions containing SP1 but no CCAAT sequence elements. (**B**) PQSSs in upstream regions containing CCAAT but no SP1 sequence elements. (**C**) PQSSs in upstream regions containing SP1 and CCAAT sequence elements. (**D**) PQSSs in upstream regions containing neither SP1 nor CCAAT sequence elements.

although having a maximum at the same place as the CCAAT elements themselves (151 elements in the −70 to −61 region) has virtually the same number (149) in the region of the peak of the SP1 consensus sequence distribution (−70 to −61 bases). The PQSSs that occur in genes without upstream CCAAT elements have a peak in the same regions as the SP1 sequence elements.

Figure 6 shows the effect of the presence of SP1 and CCAAT sequence elements on the PQSS distribution, focusing on the distribution in additional groupings of genes. PQSSs upstream of genes with upstream SP1 elements but no upstream CCAAT elements have a distribution very similar to the SP1 element and the PQSS distribution of all known Ensembl genes (Figure 6A). There are very few PQSSs in genes, which contain upstream CCAAT sequences but no SP1 elements, and their distribution is very flat (Figure 6B). Genes with both upstream SP1 and CCAAT elements have many more PQSS however not such a distinct maximum (Figure 6C) and genes with neither upstream SP1 nor CCAAT elements have very few PQSSs and a rather flat distribution (Figure 6D).

The distribution of PQSSs resembles that of regulatory motifs such as SP1 and CCAAT, although it would appear that upstream SP1 elements have a positive effect on the number of upstream PQSSs while the presence of CCAAT has a deleterious effect. Almost half of the total upstream PQSSs have the SP1 consensus sequence incorporated. Thus we can demonstrate that PQSSs linked to SP1 sequence motifs in the upstream regions is perhaps unsurprising but what is not necessarily so obvious is how dominant this effect is. It has been proposed that induction of quadruplex formation in promoter sequences by quadruplex-selective small molecules can be a viable therapeutics strategy (26). The present study provides support for this approach, and suggests that effort could be focused on those genes in which PQSSs are linked to SP1 sites, as is the case, for example, of the c-kit gene implicated in gastrointestinal cancers (15,27). A very recent report (28), in contrast, finds G-rich sequences in the first intron of many human genes, and considers that these are more likely to be PQSSs suitable for therapeutic intervention, in part because of the potential for structural polymorphism in the upstream sites. It is not clear that this would be a problem since it is likely that small molecule binding would tend to drive the equilibrium towards discrete quadruplex species. In addition, some PQSS sites such as those in the c-kit promoter (15,27), comprise isolated runs of just four G-tracts each, and are much less likely to participate in quadruplex polymorphism.

We also note that the presence of the zinc finger motif in SP1 may be significant in view of findings that the motif has been selected out from phage libraries to bind to quadruplex DNAs (29–31), and that transcription factors containing zinc fingers have been reported to bind to G-tract promoter sequences, notably the insulin promoter factor Pur-1/MAZ (32).

ACKNOWLEDGEMENTS

This work has been support by Cancer Research UK (programme grant C129/A4489). We are grateful to an anonymous reviewer for constructive comments. Funding to pay the Open Access publication charges for this article was provided by Cancer Research UK.

Conflict of interest statement. None declared.

REFERENCES

1. Gellert,M., Lipsett,M.N. and Davies,D.R. (1962) Helix formation by guanylic acid. *Proc. Natl Acad. Sci. USA*, **48**, 2013–2018.
2. 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.

3. Burge,S., Parkinson,G.P., Hazel,P., Todd,A.K. and Neidle,S. (2006) Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res.*, **34**, 5402–5415.
4. D'Antonio,L. and Bagga,P.S. (2004) Computational methods for predicting intramolecular G-Quadruplexes in nucleotide sequences. *CSB Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB 2004)*. Stanford University, CA, pp. 561–562.
5. Todd,A.K., Johnston,M. and Neidle,S. (2005) Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res.*, **33**, 2901–2907.
6. Huppert,J.L. and Balasubramanian,S. (2005) Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.*, **33**, 2908–2916.
7. 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.
8. Rawal,P., Kummarasetti,V.B.R., 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.
9. 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.
10. Eddy,J. and Maizels,N. (2006) Gene function correlates with potential for G4 DNA formation in the human genome. *Nucleic Acids Res.*, **34**, 3887–3896.
11. Huppert,J.L. and Balasubramanian,S. (2006) G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res.*, **35**, 406–413.
12. Phan,A.T., Kuryavyi,V., Gaw,H.Y. and Patel,D.J. (2005) Small-molecule interaction with a five-guanine-tract G-quadruplex structure from the human MYC promoter. *Nat. Chem. Biol.*, **1**, 167–173.
13. Phan,A.T., Kuryavyi,V., Burge,S., Neidle,S. and Patel,D.J. (2007) Structure of an unprecedented G-quadruplex scaffold in the human c-kit promoter. *J. Am. Chem. Soc.*, **129**, 4386–4392.
14. Ambrus,A., Chen,D., Dai,J., Jones,R.A. and Yang,D. (2005) Solution structure of the biologically relevant G-quadruplex element in the human c-MYC promoter. Implications for G-quadruplex stabilization. *Biochemistry*, **44**, 2048–2058.
15. 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.
16. Cogoi,S. and Xodo,L.E. (2006) G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription. *Nucleic Acids Res.*, **34**, 2536–2549.
17. De Armond,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.
18. 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.
19. Everett,R.D., Baty,D. and Chambon,P. (1983) The repeated GC-rich motifs upstream from the TATA box are important elements of the SV40 early promoter. *Nucleic Acids Res.*, **11**, 2447–2464.
20. Dynan,W.S. and Tjian,R. (1983) The promoter-specific transcription factor Sp1 binds to upstream sequences in the SV40 early promoter. *Cell*, **35**, 79–87.
21. Dynan,W.S. and Tjian,R. (1985) Control of eukaryotic messenger RNA synthesis by sequence-specific DNA-binding proteins. *Nature*, **316**, 774–778.
22. Kadonaga,J.T., Jones,K.A. and Tjian,R. (1986) Promoter-specific activation of RNA polymerase II transcription by Sp1. *Trends Biochem. Sci.*, **11**, 20–23.
23. Thiesen,H.J. and Bach,C. (1990) Target Detection Assay (TDA): a versatile procedure to determine DNA binding sites as demonstrated on SP1 protein. *Nucleic Acids Res.*, **18**, 3203–3209.
24. Hubbard,T.J.P., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F., Cutts,T. *et al.* (2007) Ensembl 2007. *Nucleic Acids Res.*, **35**, D610–D617.
25. Stabenau,A., McVicker,G., Melsopp,C., Proctor,G., Clamp,M. and Birney,E. (2004) The Ensembl core software libraries. *Genome Res.*, **14**, 929–933.
26. 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.
27. Fernando,H., Reszka,A.P., Huppert,J., Ladame,S., Rankin,S., Venkitaraman,A.R., Neidle,S. and Balasubramanian,S. (2006) A conserved quadruplex motif located in a transcription activation site of the human c-kit oncogene. *Biochemistry*, **45**, 7854–7860.
28. Eddy,J. and Maizels,N. (2008) Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes. *Nucleic Acids Res.*, doi:10.1093/nar/gkm1138.
29. Isalan,M., Patel,S.D., Balasubramanian,S. and Choo,Y. (2001) Selection of zinc fingers that bind single-stranded telomeric DNA in the G-quadruplex conformation. *Biochemistry*, **40**, 830–836.
30. Patel,S.D., Isalan,M., Gavory,G., Ladame,S., Choo,Y. and Balasubramanian,S. (2004) Inhibition of human telomerase activity by an engineered zinc finger protein that binds G-quadruplexes. *Biochemistry*, **43**, 13452–13458.
31. Ladame,S., Schouten,J.A., Roldan,J., Redman,J.E., Neidle,S. and Balasubramanian,S. (2006) Exploring the recognition of quadruplex DNA by an engineered Cys2-His2 zinc finger protein. *Biochemistry*, **45**, 1393–1399.
32. Lew,A., Rutter,W.J. and Kennedy,G.C. (2000) Unusual DNA structure of the diabetes susceptibility locus IDDM2 and its effect on transcription by the insulin promoter factor Pur-1/MAZ. *Proc. Natl Acad. Sci. USA*, **97**, 12508–12512.