Greglist: a database listing potential G-quadruplex regulated genes

Ren Zhang¹, Yan Lin² and Chun-Ting Zhang²*

¹Department of Epidemiology and Biostatistics, Tianjin Cancer Institute and Hospital, Tianjin 300060, China and ²Department of Physics, Tianjin University, Tianjin 300072, China

Received August 13, 2007; Revised September 15, 2007; Accepted September 17, 2007

ABSTRACT

The double helix is a conformation that genomic DNA usually assumes; under certain conditions, however, guanine-rich DNA sequences can form a four-stranded structure, G-quadruplex, which is found to play a role in regulating gene expression. Indeed, it has been demonstrated that the G-quadruplex formed in the c-MYC promoter suppresses its transcriptional activity. Recent studies suggest that G-quadruplex motifs (GQMs) are enriched in human gene promoters. To facilitate the research of G-quadruplex, we have constructed Greglist, a database listing potentially G-quadruplex regulated genes. Greglist harbors genes that contain promoter GQMs from genomes of various species, including humans, mice, rats and chickens. Many important genes are found to contain previously unreported promoter GQMs, such as ATM, BAD, AKT1, LEPR, UCP1, APOE, DKK1, WT1, WEE1, WNT1 and CLOCK. Furthermore, we find that not only protein coding genes, 126 human microRNAs also contain promoter GQMs. Greglist therefore provides candidates for further studying G-quadruplex functions and is freely available at http://tubic.tju.edu.cn/greglist.

INTRODUCTION

The double helix structure is a conformation that genomic DNA usually assumes; however, DNA can form other non-classical structures as well (1). For instance, under certain conditions, guanine-rich DNA sequences can form a special structure called G-quadruplex. The discovery of G-quadruplex can be traced back to G-quartets, planar arrays of four guanines held together by hydrogen bonds, which were found by Davies and coworkers (2) about 5 decades ago. Later Sen and Gilbert (3) discovered G-quadruplex, a four-stranded structure that is stabilized by G-quartets. As an example, readers may visit www.rcsb.org to view the 3-dimensional (3D) structure of a G-quadruplex (PDB code: 1XAV), which is formed in the promoter regions of the c-MYC gene (4). Sequences with high potential to form G-quadruplex have been found in many different genomic regions, suggesting diverse roles of G-quadruplexes (5–11). For instance, telomeric repeats in virtually all eukaryotes have the ability to form G-quadruplexes (10,11), offering a protection for the telomere 3’ overhang (12,13), which is essential for cell survival.

Recent interests on G-quadruplexes have been focused on its role in transcriptional regulation. By using electron microscopy, Maizels and coworkers (14) observed that the G-quadruplex structure is formed cotranscriptionally in vivo. Indeed, Hurley and coworkers have demonstrated that the region upstream of the c-MYC promoter forms a G-quadruplex, removal of which results in an increase, whereas its stabilization results in a decrease in basal transcriptional activity of this promoter, suggesting promoter G-quadruplexes as transcriptional repressor elements (15).

Sequences containing G-quadruplex motifs (GQMs) in promoter regions have only been reported for about 10 genes, including c-MYC (15–17), VEGF (18), BCL-2 (19), c-KIT (20,21) and some others (22,23). Recent bioinformatics studies, however, showed that GQMs are prevalent in the human genome (24,25). Furthermore, GQMs were found to be highly enriched in human gene promoters with more than 40% promoters containing at least 1 GQM (26).

To facilitate the study of the role of promoter G-quadruplexes, we constructed Greglist, a database listing potential G-quadruplex REgulated Genes, i.e. genes that contain promoter GQMs. The database provides detailed information about the number, the position and the sequence of promoter GQMs from genes of various species. Many important genes are found to contain previously unreported promoter GQMs, such as ATM, BAD, AKT1, LEPR, UCP1, APOE, DKK1, WT1, WEE1, WNT1 and CLOCK. Furthermore, we found that not only protein coding genes, 126 human microRNAs also contain promoter GQMs. Greglist contains

*To whom correspondence should be addressed. Tel: +86 22 27402987; Fax: +86 22 2335 8329; Email: etzhang@tju.edu.cn

© 2007 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
candidates for further studying G-quadruplex functions and is another device added to the existing online G-quadruplex toolbox.

DATABASE CONSTRUCTION AND DESCRIPTION

Greglist of the current version contains genes that have promoter GQMs in the genomes of human, mouse, rat and chicken. Table 1 provides a descriptive statistics of the content of the database. We generally defined sequences 1 kb upstream of transcription start site (TSS) as promoter regions. These sequences were downloaded from Ensembl using the software BioMart. The dataset used was Ensembl 45 and human, mouse, rat and chicken genome sequences were based on the versions of NCBI36, NCBI36, RGS3C3.4 and WASHUC2, respectively. The software Quadparser (26) was used to find the promoter GQM, which is G₃,N₁₋G₃,N₁₋G₃,N₁₋G₃, where N denotes any nucleotide. In addition, the G-quadruplex structure can be formed on either of the two DNA strands; therefore the motif of C₃,N₁₋C₃,N₁₋C₃,N₁₋C₃, was also used, which suggests the capability of the G-quadruplex formation on the complementary strand.

So far, only about 10 genes have been reported to contain promoter GQMs. In Greglist, however, a lot more genes that contain promoter GQMs are listed. For instance, these genes include ATM, BAD, AKT1, LEPR, UCP1, APOE, DKK1, WT1, WEE1, WNT1, CLOCK, ATFI and BMP2, which have critical functions in various cellular processes, such as apoptosis and transcriptional regulation. Table 2 lists a sample of 30 genes that contain promoter GQMs with the position of GQMs and gene functions.

In addition, we found that not only protein coding genes, many microRNAs, such as hsa-mir-639 and hsa-mir-381, also contain promoter GQMs. Totally 126 human microRNAs were found to have promoter GQMs. To get a full list of these microRNAs, refer to the Supplementary Table. MicroRNAs have emerged as important regulators of gene expression. The finding that promoter regions of microRNA genes contain GQMs necessitates further studies to address the role of G-quadruplexes in microRNA regulation.

Of note, the presence of a GQM only suggests the potential of a sequence to form G-quadruplex. In addition, the G-quadruplex structure is a dynamic structure that is formed upon denaturation of the DNA duplex. Therefore caution must be taken to interpret the data in Greglist. In other words, gene records in Greglist provide a starting point for further analysis of the potential G-quadruplex structure in these genes. Furthermore, Huppert et al. (26) reported that more than 40% of human genes contain promoter GQMs, however, in Greglist, ~32% human genes do. This is likely because in Ref. (26), only less than 20 000 known genes were used, whereas in the current study, more than 30 000 human genes, including those classified as novel and those encode RNAs were included. Therefore, Greglist is made to be inclusive, not exclusive.

Gene names, Ensembl IDs, RefSeq IDs, numbers of GQMs, distance of the GQM to TSS, functional description of gene ontology, sequences containing the GQM and coding sequences of the gene, were extracted from Ensembl database and Quadparser output files. All the data were then organized by using an open-source management system, MySQL, which allows rapid data retrieval. All gene records have been linked directly to corresponding entries in Ensembl. Users can browse each entry or download all records. Because of the large volume of data, a good searching function is important for this database. In Greglist, users can perform searches by inputting gene accession numbers or names at the homepage, and then click ‘Go’. To perform more detailed searches, users can click ‘Search’, and then in the new page, more detailed searching options are provided. For instance, users can search by gene ontology terms to get a list of genes that have desired functions. To further facilitate searching the gene of interest, we installed Blast program locally. So users can input the coding sequence of their gene of interest and perform Blast searches to find homologous ones.

Many online resources for G-quadruplexes are available. These include G4P calculator (14), QGRS Mapper (27), Quadfinder (28), which are online programs or web servers for predicting G-quadruplexes. GRSDB (29) is a database of quadruplex forming G-rich sequences in alternatively processed mammalian pre-mRNA sequences. Greglist is another device added to the existing online G-quadruplex toolbox.

We plan to include more species in future versions of Greglist. In addition, with the availability of more experimental data, we plan to integrate experimental evidence in corresponding entries. Furthermore, although the GQM used in Quadparser is quite commonly used, there are other motifs that have potential to form G-quadruplexes, and we also plan to include these motifs in future versions of the database. We welcome
users’ comments, corrections and new information, which will be used for updating.

Greglist is freely available at the website: http://tubic.tju.edu.cn/greglist, and should be cited with the present publication as reference.

SUPPLEMENTARY DATA
Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS
We would like to thank Prof. Jingchu Luo and Dr Ge Gao for their helps. We also thank two anonymous referees for their constructive comments. The present work was supported in part by NNSF of China (Grant No. 90408028). Funding to pay the Open Access publication charges for this article was provided by NNSF of China (Grant No. 90408028).

Conflict of interest statement. None declared.

Table 2. A list of 30 human genes that have not been previously reported to contain promoter G-quadruplex motifs

| No. | Abbreviation | Gene name | Ensembl ID | Function or associated disease | Reference | Number of GQM | Distance to TSS |
|-----|--------------|-----------|------------|--------------------------------|-----------|---------------|----------------|
| 1   | WNT1         | Wingless-type MMTV integration site family, member 1 | ENSG00000125084 | The Wnt signaling pathway, CNS development | (30)      | 1             | 193            |
| 2   | WNT5A        | Wingless-type MMTV integration site family, member 5A | ENSG00000114251 | The Wnt signaling pathway, vertebrate development | (31)      | 2             | 567, 936       |
| 3   | LEPR         | LEPTIN receptor | ENSG00000116678 | Energy metabolism | (32)      | 3             | 310, 372, 495  |
| 4   | UCP1         | Uncoupling protein 1 | ENSG00000109424 | Energy metabolism | (33)      | 2             | 89, 224        |
| 5   | APOE         | Apolipoprotein E | ENSG00000130203 | Alzheimer’s disease | (34)      | 4             | 46, 65, 407, 739 |
| 6   | ATM          | Ataxia telangiectasia mutated | ENSG00000149311 | Ataxia telangiectasia | (35)      | 1             | 59             |
| 7   | PAX8         | Paired box gene 8 | ENSG00000125618 | Permanent congenital hypothyroidism | (36)      | 1             | 133            |
| 8   | SOX1         | SRY (sex determining region Y)-box 1 | ENSG00000203883 | Lens development | (37)      | 3             | 80, 726, 826   |
| 9   | SOX10        | SRY (sex determining region Y)-box 10 | ENSG00000100146 | Waardenburg–Hirschsprung disease | (38)      | 2             | 130, 313       |
| 10  | HDAC1        | Histone deacetylase 1 | ENSG00000116478 | Histone modification | (39)      | 1             | 34             |
| 11  | TGFβ1        | Transforming growth factor, beta 1 | ENSG00000105329 | TGFβ signaling | (40)      | 1             | 151            |
| 12  | SMAD2        | MAD homolog 2 | ENSG00000175387 | TGFβ signaling | (41)      | 2             | 235, 450       |
| 13  | DKK1         | Dickkopf homolog 1 | ENSG00000107984 | TGFβ signaling | (42)      | 1             | 136            |
| 14  | CLOCK        | Clock homolog | ENSG00000134852 | Circadian rhythms | (43)      | 3             | 147, 341, 692  |
| 15  | WEE1         | WEE1 homolog | ENSG00000166483 | Cell cycle control | (44)      | 1             | 542            |
| 16  | BAD          | BCL2-antagonist of cell death | ENSG0000002330 | Apoptosis | (45)      | 3             | 116, 628, 756  |
| 17  | AKT1         | V-akt murine thymoma viral oncogene homolog 1 | ENSG00000142208 | Apoptosis | (46)      | 1             | 61             |
| 18  | GATA4        | GATA-binding protein 4 | ENSG00000136574 | Heart development | (47)      | 1             | 314            |
| 19  | MYOD1        | Myogenic differentiation 1 | ENSG00000129152 | Muscle development | (48)      | 2             | 128, 216       |
| 20  | WT1          | Wilms tumor 1 | ENSG00000184937 | Kidney development | (49)      | 2             | 168, 900       |
| 21  | GDF1         | Growth differentiation factor 1 | ENSG00000135414 | Left-right patterning | (50)      | 4             | 78, 166, 327, 766 |
| 22  | BMP2         | Bone morphogenetic protein 2 | ENSG00000125845 | Bone development | (51)      | 1             | 163            |
| 23  | MEF2D        | MADS box transcription enhancer factor 2D | ENSG00000116604 | Heart development | (52)      | 4             | 18, 85, 169, 232 |
| 24  | STAT6        | Signal transducer and activator of transcription 6 | ENSG00000166888 | Immunity | (53)      | 1             | 505            |
| 25  | SOCS1        | Suppressor of cytokine signaling 1 | ENSG00000185338 | Immunity | (54)      | 5             | 112, 211, 534, 578, 758 |
| 26  | MMP2         | Matrix metalloproteinase 2 | ENSG00000167346 | Function of extracellular matrix | (55)      | 1             | 576            |
| 27  | MAPK2        | Mitogen-activated protein kinase 2 | ENSG00000162889 | MAP kinase pathway | (56)      | 2             | 100, 137       |
| 28  | ATF1         | Activating transcription factor 1 | ENSG00000123268 | Transcriptional regulation | (57)      | 1             | 36             |
| 29  | TAF2         | TAF2 RNA polymerase II | ENSG00000064313 | Transcriptional regulation | (58)      | 1             | 296            |
| 30  | RING1        | Ring finger protein 1 | ENSG00000204227 | Transcriptional regulation | (59)      | 4             | 501, 559, 677, 938 |

D374   Nucleic Acids Research, 2008, Vol. 36, Database issue

Nucleic Acids Research, 2008, Vol. 36, Database issue
REFERENCES

1. Perez-Martijn, J. and de Lorenzo, Y. (1997) Clues and consequences of DNA binding in transcription. *Ann. Rev. Microbiol.*, **51**, 593–628.

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

3. Corden, J.D. and Gilman, A. (1988) Formation of parallel-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. *Nature*, **334**, 364–366.

4. Ambrus, A., Chen, D., Carver, M., Ambrus, A., and Hurley, L.H. (2005) Formation of parallel-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. *Nature*, **334**, 364–366.

5. Sun, D., Guo, K., Rusche, J.J. and Hurley, L.H. (2005) Solution structure of the biologically relevant G-quadruplex element in the human c-MYC promoter. Implications for G-quadruplex stabilization. *Biochemistry*, **44**, 2050–2058.

6. Simonsson, T. (2001) G-quadruplex DNA structures—variations on a theme. *Biol. Chem.*, **382**, 621–628.

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

8. Maizels, N. (2006) Dynamic roles for G4 DNA in the biology of eukaryotic cells. *Nat. Struct. Mol. Biol.*, **13**, 1055–1059.

9. Han, H. and Hurley, L.H. (2000) G-quadruplex DNA: a potential target for anti-cancer drug design. *Trends Pharmacol. Sci.*, **21**, 136–142.

10. Burge, S., Rankin, S., Lapi, P., Todd, A.K. and Neidle, S. (2001) G-quadruplex DNA structures—variations on a theme. *Biochemistry*, **34**, 5402–5415.

11. Blackburn, E.H. (1991) Structure and function of telomeres. *Nature*, **350**, 569–573.

12. Williamson, J.R. (1994) G-quartet structures in telomeric DNA. *Annu. Rev. Biophys. Biomol. Struct.*, **23**, 703–730.

13. Huppert, J.L. and Balasubramanian, S. (2005) Prevalence of G-quadruplexes in the human genome. *Nucleic Acids Res.*, **33**, 2901–2907.

14. Todd, A.K., Johnston, M. and Neidle, S. (2005) Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res.*, **33**, 2901–2907.

15. Siddiqui-Jain, A., Grand, C.L., Bearss, D.J. and Hurley, L.H. (2002) Solution structure of the biologically relevant G-quadruplex element in the proximal promoter region of the human VEGF gene. *Proc. Natl Acad. Sci. USA.*, **99**, 11593–11598.

16. Simonsen, T., Pecinka, P. and Kubista, M. (1998) DNA tetraplex formation of guanine tetrads between hairpin loops. *Nature*, **342**, 825–829.

17. Duquette, M.L., Haila, P., Vincent-Ja, T., Kenn,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.

18. 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.

19. Grand, C.L., Han, H., Munoz, R.M., Weitman, S., Von Hoff, D.D., Hurley, L.H. and Bearss, D.J. (2002) The cationic porphyrin TMPyP4 down-regulates c-MYC and human telomerase reverse transcriptase expression and inhibits tumor growth in vivo. *Mol. Cancer Ther.*, **1**, 565–573.

20. Sun, D., Guo, K., Rusche, J.J. and Hurley, L.H. (2005) Facilitation of a structural transition in the polypurine/polypyrimidine tract within the proximal promoter region of the human VEGF gene by the presence of potassium and G-quadruplex-interactive agents. *Nucleic Acids Res.*, **33**, 6070–6080.

21. Dai, J., Desheimer, T.S., Chen, D., Carver, M., Ambrus, A., Jones, R.A. and Yang, D. (2006) An intramolecular G-quadruplex structure with mixed parallel/antiparallel G-strands formed in the human BCL-2 promoter region in solution. *J. Am. Chem. Soc.*, **128**, 1096–1098.

22. Rankin, S., Reszka, A.P., Huppert, J., Zloch, 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.

23. 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.

24. Huang, R.M., Woodford, K.J., Weitzmann, M.N. and Usdin, K. (1996) The chicken beta-globin gene promoter forms a novel "inchred" tetrahedral structure. *J. Biol. Chem.*, **271**, 5208–5214.
44. Russell, P. and Nurse, P. (1987) Negative regulation of mitosis by wee1+, a gene encoding a protein kinase homolog. Cell, 49, 559–567.
45. Datta, S.R., Dudek, H., Tao, X., Masters, S., Fu, H., Gotoh, Y. and Greenberg, M.E. (1997) Akt phosphorylation of BAD couples survival signals to the cell-intrinsic death machinery. Cell, 91, 231–241.
46. Nicholson, K.M. and Anderson, N.G. (2002) The protein kinase B/Akt signalling pathway in human malignancy. Cell. Signal., 14, 381–395.
47. Molkentin, J.D., Lin, Q., Duncan, S.A. and Olson, E.N. (1997) Requirement of the transcription factor GATA4 for heart tube formation and ventral morphogenesis. Genes Dev., 11, 1061–1072.
48. Takeda, K., Tanaka, T., Shi, W., Matsumoto, M., Minami, M., Kashiwamura, S., Nakanishi, K., Yoshida, N., Kishimoto, T. and Akira, S. (1996) Essential role of Stat6 in IL-4 signalling. Nature, 380, 627–630.
49. Alexander, W.S., Starr, R., Fenner, J.E., Scott, C.L., Handman, E., Sprigg, N.S., Corbin, J.E., Cornish, A.L., Darwiche, R. et al. (1999) SOCS1 is a critical inhibitor of interferon gamma signaling and prevents the potentially fatal neonatal actions of this cytokine. Cell, 98, 597–608.
50. Giannelli, G., Falk-Marzillier, J., Schiraldi, O., Stetler-Stevenson, W.G. and Quantra, V. (1997) Induction of cell migration by matrix metalloprotease-2 cleavage of laminin-5. Science, 277, 225–228.
51. Hogan, B.L. (1996) Bone morphogenetic proteins: multifunctional regulators of vertebrate development. Genes Dev., 10, 1580–1594.
52. Edmondson, D.G., Lyons, G.E., Martin, J.F. and Olson, E.N. (1994) Mef2 gene expression marks the cardiac and skeletal muscle lineages during mouse embryogenesis. Development, 120, 1251–1263.