In silico comparative analysis of promoters derived from plant pararetroviruses

Dipinte Gupta¹ · Rajiv Ranjan¹

Received: 7 September 2017 / Accepted: 8 November 2017 / Published online: 28 November 2017
© Indian Virological Society 2017

Abstract Promoters are specific sequence of nucleotides present upstream of gene coding region involved in initiation and regulation of transcription. Multiple cis-acting element forms the architecture of promoter to which trans-acting nucleic binding factors bind and regulates its activity. Since 1980 genome of pararetrovirus, are being exploited for developing efficient promoters. Among all of them Cauliflower mosaic virus is the most widely used promoter for gene expression. The basic rationale behind the strength of promoter lies in the sequence of cis elements and the spacer nucleotide elements between them, thereby strength of these promoter fragments can be regulated by altering these nucleotide sequences. In the present study sequence of eight putative promoters of plant pararetrovirus are retrieved from National Centre for Biotechnology Information (NCBI) website. These sequence are subjected to various bioinformatics tools comprises of Clustal W, Plant Care, MathInspector, ModelInspector for establishing the phylogenetic similarity, to identify the quantity and quality of present cis-elements, to find the various common transcription factors binding sites and to determine the presence of module for various specific functions respectively. A range of 28.80–56.0 percentage identification was observed in phylogenetic analysis, with the greatest similarity in Mirabilis mosaic virus and Dahlia mosaic virus. A

broad range of cis-elements (51), transcription factor binding site (512) was obtained and 60% observed module are in combination with DOFF motif which shows a function relevance with abiotic stress inducibility. The present study had revealed the functional significance of these elements in gene regulation of pararetrovirus genome and also gives a overall idea for designing novel synthetic promoter.

Keywords Pararetrovirus · TFBS · Cis-element · In silico

Introduction

Pararetrovirus is a family of viruses which is found in both plants and animals. These are classified into *Hepadnaviridae* (infects animals) and *Caulimoviridae* (infects plants) [23]. These viruses encapsulate a double stranded (ds) DNA circular genome and replicate through an RNA intermediate. They do not integrate into host genome either they deposit in the form of episomal bodies [2]. Two major transcriptional promoters are found in this group, one which transcribe the full length transcript (-Flt-) and other which transcribes sub-genomically (-Sgt-), Flt and Sgt promoters are located at 3' terminus of ORF V and VI respectively [32]. Since 90s promoter from family *caulimoviridae* of plant pararetrovirus has proved its suitability in transcriptional regulation [26]. Scientist had described different type of Caulimovirus, their members, are exemplified by *Cauliflower mosaic virus* (CaMV) [11], *Figwort mosaic virus* (FMV) [25], *Peanut chlorotic streak virus* (PCISV) [24], *Cassava vein mosaic virus* (CVMV) [3], *Strawberry vein banding virus* (SVBV) [20], *Mirabilis mosaic virus* (MMV) [6], *Dahlia mosaic virus* (DMV) [16] and *Rice tungro bacilliform virus* (RTBV) [31].

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s13337-017-0410-8>) contains supplementary material, which is available to authorized users.

✉ Rajiv Ranjan
rajivranjanbt@gmail.com

¹ Plant Biotechnology Lab, Department of Botany, Faculty of Science, Dayalbagh Educational Institute (Deemed University), Dayalbagh, Agra 282005, India

Detail study reported on eukaryotic promoter had revealed that transcriptional activity depends on combinatorial and synergistic interaction of multiple cis-acting elements and trans-acting nuclear binding factor [8]. Recognition and understanding of these elements were gained by doing several experiments comprises of deletions, mutagenesis, linker scanning analysis and element relocation chromatin immunoprecipitation [9]. Two nuclear binding protein factors, known as activating sequence factor – 1 and – 2 (ASF-1 and ASF-2) from tobacco have been well characterized [17]. TATA box, box at 25–30 bases, CCAAT at – 80 from transcription start site and other regulatory elements deliberate greater and regulatory activity on a promoter. Cis-elements establish the polarity to which trans-acting nuclear binding factors binds [14]. Alteration in sequences, orientation or position can dramatically increase or reduces rate of protein synthesis [27]. Sequencing and annotation of a large number of eukaryotic genomes has made available an enormous amount of information regarding genetic coding sequences (CDS) [5]. These data can be effectively utilized for studying and modifying the expression of genes if the location and contribution of cis-regulatory regions, which control spatial and temporal regulation of gene expression, are available. Specific sequence elements in DNA regulate the transcription by recruiting transcription factors transcript level [12] as in the case of *rbcS-3A*, suggesting critical sequence requirements for the activation of gene expression. [7].

In the present study, detailed analysis of cis-elements and transcription factors of eight promoters of plant pararetrovirus were performed by employing bioinformatics tools. Promoter of CaMV, CVMV, DMV, FMV, PC1SV, RTBV, MMV and SVBV were used. This is the first report where promoters of plant pararetrovirus were analyzed for its functional motif. Our study represents a wide comparative data which would be helpful for designing synthetic promoter from pararetrovirus.

Material and method

Data mining

Eight promoters conscripted as CaMV, CMV, DMV, FMV, MMV, PC1SV, RTBV and SVBV were retrieved from NCBI website having web address <https://www.ncbi.nlm.nih.gov/>. The length for each promoter was taken from the scientific data reported for maximum expression. Following NCBI accession ID's are used for sequence retrieval JQ280535.1 (CaMV), NC_001648.1 (CVMV), JX272320.1 (DMV), NC_003554.1 (FMV), AF454635.1 (MMV), U13988 (PC1SV), AJ314596 (RTBV), X97304.1 (SVBV).

GC content and free energy of the promoter was calculated with help of the web based software Oligo calculator (<http://biotools.nubic.northwestern.edu/OligoCalc.html>) [15].

Phylogenetic analysis

Promoter sequences of eight putative promoters were aligned by using EMBL software Clustal Omega (<http://www.ebi.ac.uk/Tools/msa/clustalo/>). The aligned sequences were used for construction of neighbor-joining phylogenetic tree by using Molecular Evolution Genetic Analysis 6 (MEGA6) tool [30].

Detailed description of cis-elements

Expression of promoters depends on the quality and quantity of cis-elements. In order to study the cis-elements present in promoters, each promoter was searched for presence of cis-elements by using the database PLANT CARE (<http://bioinformatics.psb.ugent.be/webtools/plantcare/html/>) [18]. Both “+” and “–” strand were analyzed for the presence of cis-elements. A combined table was assembled, showing the presence of various cis elements in all eight promoter sequences.

Comparative analysis of transcription factor binding site

Fasta format of sequences derived from eight pararetrovirus promoters was subjected to Mat-Inspector (<https://www.genomatix.de/cgi-bin/matinspector>) [4] for inspecting TFBS. General Core promoter elements and plants matrix group were selected with 0.75 core similarity and Optimized matrix similarity. These sequences were searched for common transcription factors among them by using common TFs search (<https://www.genomatix.de/cgi-bin/gems/launch.pl?s=6bfe9eebfacd4bd991d19e7681ab0689;GEMS=1;TASK=commonTFs>).

Module and framework study

All the retrieved promoter sequences were subjected to Model Inspector for determining the promoter modules. Latest library module 6.1 with plant module as a model group was selected. A range of 100–1000 were kept for searching the modules in both positive and negative strand with a minimum number of two elements in each module. A framework was also tried to establish by using web based software (<https://www.genomatix.de/cgibin/gems/launch.pl?s=6bfe9eebfacd4bd991d19e7681ab0689;GEMS=1;TASK=framework>) [10].

Result

Data mining

A range of 250–714 bp was observed for the promoter sequences with an average GC content of 41.75. In our study we had observed the enhancement in free energy with increase in base pair of sequences. All the sequences with their respective number of bases, GC content and free energy had been provided in supplementary table 1.

Phylogenetic analysis

A percentage identification table, developed from aligned promoter fragment had shown a maximum of 56.0 similarities between MMV and DMV. Phylogenetically, the promoters can be divided into two clades among which clade 1 can be subdivided into two sub-clade 1a and 1b among which sub-clade 1a can be further divided into two sub-sub clade 1a having SVBV and PCSV on one branch and CVMV on another branch while sub-sub clade b1 have MMV and DMV. Whereas sub-clade 1b has CaMV. Clade 2 has RTBV and FMV on two separate branches (Fig. 1).

Detailed description of cis-elements

A search for the presence of the cis-elements in their nucleotide sequence of all eight promoters had shown the diverse pattern of occurrence. A total of 51 cis elements had been found in the promoter sequences used in this piece of work (supplementary table 2). TATA Box and CAAT Box are the most conserved cis elements presence throughout the family pararetrovirus with highest copy number,

Fig. 1 Cladogram of pararetroviruses based on promoter sequences by using MEGA 6 through neighbor's joining method

CGTCA, cis-acting regulatory element involved in the methyl jasmonate (MeJA) responsiveness was found in seven searched promoter sequences except in SVBV. Similarly Skn1_Motif Cis acting regulatory element required for endosperm expression is present in all except DMV. GATA, I box, TCA elements, TGACG, Unnamed 4 are the cis-elements which was found in more than 62.5% of searched promoters while rest of the cis elements are found in less than 62.5 searched promoters nucleotide sequences (Fig. 2).

Comparative analysis of transcription factor binding site

The promoter sequences were analyzed for the available transcription factor binding site. A total of 521 matches were found in all sequences of total 3521 bp. A match of 59, 76, 81, 38, 65, 53, 101, 48 TFBS was found in CaMV, CVMV, DMV, FMV, RTBV, PC1SV, MMV, SVBV respectively. Among 521 TFBS, 288 were found on “+” strand while rest on “–” strand. Detailed description of TFBS among all the accession is given in supplementary table 5 and a graphical representation is given in Fig. 3. Common TFBS investigation was also performed which revealed 84 common matches among the sequences. Highest frequency was found of the P\$NACF matrix family with a 27 common matches having the *p* value 0.0204163. Transcription factors from this family matrix are common to all eight sequences with 8, 1, 6, 1, 3, 4, 6, 3 copy numbers in SVBV, PC1SV, MMV, FMV, CVMV, CaMV and DMV respectively. A detailed description of

Fig. 2 Cis-elements common to promoters. Graphical representation of most conserved cis-element present in the promoters (Y-axis represents the number of cis-element present, X-axis represents name of cis-element present)

Fig. 3 Graphical representation of transcription factor binding site present in searched promoter sequences

common TFBS having the total matches of 10 and above are given in Supplementary table 3.

Module and framework study

In silico functional analysis using ModelInspector revealed 10 Module in four sequences. 4, 1, 3, 2 modules were found in PC1SV, CaMV, CVMV, and DMV respectively. Out of the 10 module found, 6 are in combination with DOFF motif. Modules are equally distributed on both “+” and “−” strand of the sequence. AHBP_DOFF_01 module comprises of *Arabidopsis* homeobox protein and DNA binding with one finger (DOF) was found in PC1SV promoter sequence. DOFF_OPAQ_03 module having P\$DOFF (DNA binding with one finger (DOF) and P\$OPAQ (Opaque-2 like transcriptional activators) were found in CaMV sequences. GBOX_GBOX_01 were found in CVMV. Two MYCL_GCCF_01 having were found in DMV. OPAQ_DOFF_01 having P\$OPAQ (Opaque-2 like transcriptional activators) P\$DOFF (DNA binding with one finger (DOF) was found in PC1SV. OCSE_DOFF_01 and GTBX_GTBX_01 are the modules found individually in CVMV. MYCL_GCCF_01 and TALE_TALE_01 are the modules found individually in DMV Promoter. Detailed description of each found module is given in supplementary table 4. An attempt was also made to study the framework of mentioned sequences but no significant results were found.

Discussion

Since three decades, promoters from pararetrovirus are proving their diverse applicability in gene regulation. Promoters have several conserved motifs, which plays the essential role in its regulation [21]. In our study we had found the maximum homology between MMV and DMV which signifies that maximum similar conserved motifs were present in MMV and DMV. MMV and DMV promoter was found much more efficient then CaMV [6, 27]. A comparative analysis was done to examine the efficiency of plant pararetrovirus promoter which includes chimeric promoters made with duplicated enhancer sequences from the *Cauliflower mosaic virus* (D35S), the *Cassava vein mosaic virus* (DCsVMV), the *Figwort mosaic virus* (DFMV), the *Mirabilis mosaic virus* (DMMV) and the *Peanut chlorotic streak virus* (DPCLSV); their finding suggests that DMMV promoter is able to direct gene expression in citrus as strongly as the D35S promoter and represents great application potential in citrus biotechnology [7]. All eight promoters have little sequence homology except for some small regulatory motifs. They can therefore be used for expression of multiple genes in transgenic plants. Thereby strong heterologous promoter can be developed using these promoters.

Minimal promoter alone have weaker transcriptional activity, thereby cis-elements, to which tissue, development-specific transcription factors bind, individually or in combinations, usually determine the spatio-temporal expression pattern of a promoter at the transcriptional level [1]. The basic rationale behind the strength of promoter lies in the notation of the cis-acting elements. In present study an average range of 20 to 25 cis elements in a single effective promoter. CAATT box, a common cis-acting element in promoter and enhancer regions and TATA Box, a core promoter elements are found in highest copy number which indicates that they are most conserved cis-acting elements of promoter. TGACG an 'as-1' like enhancer element is found to be present in maximum promoters which includes CaMV, DMV, PC1SV, CVMV, MMV and FMV. According to scientific study Electrophoretic mobility shift assay (EMSA) and super shift analysis confirmed the binding of tobacco transcription factor TGA1a to the enhancer region of the DaMVFLt4- promoter. *DNaseI* footprinting had shown that TGA1a specially interacted with the activation sequence-1 (as-1) of DaMVFLt4- promoter [27]. Presence of other conserved cis-elements such as GATA, I box which plays a role of a light responsive element, TCA elements, a cis acting element involved in salicylic acid responsiveness, Unnamed 4 had shown their vital role for promoter development and gene regulation.

MatInspector revealed considerable diversity in TFBSs, revealing their putative roles in various plant processes. A wide range of transcription factors diversity is observed in the present study. P\$AHBP is the most abundant element reportedly involved in embryo, shoot, root patterning, shade growth control, organ fate and stem cell proliferation of an *arabidopsis* [19]. DMV possess the highest copy number of this element and SVBV have second highest copy number of this *arabidopsis* homeobox protein which shows the functional importance of this family matrix transcription factors. O\$VTBP, TATA Binding Protein Factor (TBP) is the second most abundant matrix family found in or promoter analysis. TBP is a general transcription factor that binds to the TATA box of a gene's promoter [28] and is crucial for promoter activity equally in plants, animals and viruses in gene-specific expression [13, 29]. These observations suggested the evolutionary importance of these motifs in which they are conserved by nature and presence in accordance with their demanding function.

Viruses use the host machinery for its replication, although pararetrovirus does not integrate its DNA into host genome. Infection of a virus is supposed to generate the physiological stress conditions in the plant system; Thereby a hypothesis can be drawn stating the promoters of virus should have some motif which gives its strength to remain active in adverse conditions. Doff motif is

associated with stress inducibility [22]. In our study 60% module were in combination with Doff motif, which strongly reflects the close relationship of stress relating Doff motif and promoters from virus genome.

In silico identification of cis-elements/transcription factor is an important step to understand gene expression and regulation. The presence of various cis-elements in form of function module in different pararetrovirus promoter provides evidence for their significant involvement in ectopic gene expression. Using generated data promoters can be predicted to very high level of certainty which could pave the way for their use in plant biotechnology.

Acknowledgements We greatly acknowledge DAE-BRNS (Mumbai), India (Project Sanction No. 37(1)/14/40/2014- BRNS/1423) for providing financial assistance.

References

1. Benfey PN, Ren L, Chua NH. Combinatorial and synergistic properties of CaMV 35S enhancer subdomains. *EMBO J*. 1990;9:1685–96.
2. Brunt AA, Kitajima EW. Intracellular location and some properties of Mirabilis mosaic virus, a new member of the Cauliflower mosaic group of viruses. *Phytopath Z*. 1973;76:265–75.
3. Calvert L, Ospina M, Shepherd RJ. Characterization of cassava vein mosaic virus: a distinct plant pararetrovirus. *J Gen Virol*. 1995;76:1271–6.
4. Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, Klingenhoff A, Frisch M, Bayerlein M, Werner T. MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*. 2005;21(13):2933–42.
5. Czuee M, Mookherjee S, Rajasekaran G, Bansa M. DNA free energy-based promoter prediction and comparative analysis of arabidopsis and rice genomes. *Plant Physiol*. 2011;156:1300–11.
6. Dey N, Maiti LB. Structure and promoter/leader deletion analysis of mirabilis mosaic virus (MMV) full length transcript promoter in transgenic plants. *Plant Mol Biol*. 1999;40:771–82.
7. Dutt M, Erpen L, Ananthakrishnan G, Barthe GA, Brlansky RH, Maiti IB, Grosser JW. Comparative expression analysis of five caulimovirus promoters in citrus. *Plant Cell Tiss Organ Cult*. 2016;126:229–38. <https://doi.org/10.1007/s11240-016-0993-6>.
8. Fang RX, Nagy F, Sivasubramaniam S, Chua NH. Multiple cis regulatory elements for maximal expression of the Cauliflower mosaic virus 35S promoter in transgenic plants. *Plant Cell*. 1989;1:142–50.
9. Farnham PJ. Insights from genomic profiling of transcription factors. *Nat Rev Genet*. 2009;10:605–16.
10. Frech K, Danescu-Mayer J, Werner T. A novel method to develop highly specific models for regulatory units detects a new LTR in GenBank which contains a functional promoter. *J Mol Biol*. 1997;270(5):674–87.
11. Gardner C, Howarth A, Hahn P, Leudi MB, Shepherd RJ, Messing J. The complete nucleotide sequence of an infectious clone of Cauliflower mosaic virus by M13mp7 shotgun sequencing. *Nucleic Acids Res*. 1981;9:2871–88.
12. Gilmartin PM, Chua NH. Spacing between GT-1 binding sites within a light responsive element is critical for transcriptional activity. *Plant Cell*. 1990;2:447–55.

13. Ilyas M, Muhammad S, Naqvi S, Mahmood T. *In silico* analysis of transcription factor binding sites in promoters of germin-like protein genes in rice. *Arch Biol Sci*. 2016;68(4):863–76. <https://doi.org/10.2298/ABS151116076I>.
14. Jadhao KR, Samal KC, Pradhan SK, Rout GR. Studies on molecular characterization of DREB gene in indica rice (*Oryza sativa* L.). *Hereditary Genet*. 2014;3(1):1000133.
15. Kibbe WA. OligoCalc: an online oligonucleotide properties calculator. *Nucleic Acids Res*. 2007;35:W43–6.
16. Kuluev BR, Knyazev AV, Chemeris AV. Activity of promoters of carnation etched ring virus and dahlia mosaic virus in tobacco protoplasts and transgenic plants. *Russ J Plant Physiol*. 2008;55(5):687.
17. Lam E. Analysis of tissue specific elements in the CaMV35S promoter. In: Nover L, editor. *Results and problems in cell differentiation: plant promoter and transcription factors*, vol. 20. Berlin: Springer; 1994. p. 181–96.
18. Lescot Magali, Dhais Patrice, Thijs Gert, Marchal Kathleen, Moreau Yves, Van de Peer Yves, Rouz Pierre, Rombauts Stephane. PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucleic Acids Res*. 2002;30(1):325–7.
19. Lohmann JU, Hong RL, Hobe M, Busch MA, Parcy F, Simon R, Weigel D. A molecular link between stem cell regulation and floral patterning in Arabidopsis. *Cell*. 2001;105(6):793–803.
20. Petrzik K, Benes V, Mraz L, Honestslegrova-Franova J, Ansoerge W, Spak J. Strawberry vein banding virus- definitive member of the genus Caulimovirus. *Virus Genes*. 1998;16:303–5.
21. Porto MS, Pinheiro MPN, Batisy VGL, Santos RCC, Filho PDAM, Lima LMD. Plant promoters: an approach of structure and function. *Mol Biotechnol*. 2014;56:38–49. <https://doi.org/10.1007/s12033-013-9713-1>.
22. Ranjan R, Dey N. Development of vascular tissue and stress inducible hybrid-synthetic promoters through Dof-1 motifs rearrangement. *Cell Biochem Biophys*. 2012;63:235–45. <https://doi.org/10.1007/s12013-012-9359-9>.
23. Ranjan R, Gupta D. *Introduction to plant promoter*. Saarbrücken: Lambert Academic Publication; 2016. p. 149.
24. Richins R, Richard D. Organization and expression of the peanut chlorotic streak virus genome. Ph.D. Dissertation, 1993, University of Kentucky, Lexington KY (for the PCISV genomic sequence: DNA EMBL Data Library GenBank Accession Number U13988).
25. Richins RD. The complete nucleotide sequence of peanut chlorotic streak virus (PCISV) DNA. EMBL Data Library GenBank Accession No. U13988. 1994.
26. Richins RD, Shepherd RJ. Physical maps of the genome of Dahlia mosaic virus and Mirabilis mosaic virus—two members of the caulimovirus group. *Virology*. 1983;124:208–14.
27. Rushton PJ, Anja R, Lipka V, Lippok B, Somssich IE. Synthetic plant promoters containing defined regulatory elements provide novel insights into pathogen- and wound-induced signaling. *Plant Cell*. 2002;14:749–62.
28. Russell PJ. *iGenetics: a molecular approach*. 3rd ed. London: Pearson Education; 2009.
29. Sriraman V, Bartsch JW, Rittger A, Mulders SM, Richards JS. Regulated expression of ADAM8 (a disintegrin and metalloprotease domain 8) in the mouse ovary: evidence for a regulatory role of luteinizing hormone, progesterone receptor, and epidermal growth factor-like growth factors. *Biol Reprod*. 2008;78(6):1038–48.
30. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol*. 2013;30(12):2725–9.
31. Yin Y, Beachy RN. The regulatory region of the Rice tungro bacilliform virus promoters and interacting nuclear factors in rice (*Oryza sativa* L.). *Plant J*. 1996;7(6):969–80.
32. Zhan X, Haley A, Richardson K, Morris B. Analysis of the potential promoter sequences of African cassava mosaic virus by transient expression of the fl-glucuronidase gene. *J Gen Virol*. 1991;72:2849–52.