Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Identification and analysis of putative promoter motifs in *Flavivirus* genome

Pallavi Somvanshi[1,] *, Vijai Singh[2] and Prahlad Kishore Seth[1]

[1]Bioinformatics Centre, Biotech Park, Lucknow; [2]Aquatic Microbes Section, National Bureau of Fish Genetic Resources, Canal Ring Road, P.O. Dilkusha, Lucknow; Pallavi Somvanshi* - Email: psomvanshi@gmail.com; ***** Corresponding author

received August 07, 2008; revised October 15, 2008; accepted October 30, 2008; published December 06, 2008

Abstract:
The genus *Flavivirus* comprises medically significant pathogenic virus; causing several infections in humans worldwide. *Flavivirus* genomes are 10-11 kb approximately and encode both structural and non structural region. The non structural region plays fundamental role in the stability, regulation and cell cycle of virus. The complete genomes of 26 *Flavivirus* were used for identification of promoter motifs through *in silico* approaches. The promoter sequences were encoded in merely 16 viruses and 10 viruses could not encode it. All these *in silico* identified promoter motifs was confirmed and verified with the known experimental data. This analysis suggests that presence of promoter may play a crucial role in the pattern of gene expression, regulation networks, cell specificity and development. It may also be useful for designing efficient expression vector and target specific delivery system in the gene therapy.

Keywords: *Flavivirus;* disease; promoter; regulation; cell cycle

Background:
Flavivirus consist of globally distributed group of arbovirus transmitted mainly by tick or mosquito vectors. The most significant *Flavivirus* are mosquito transmitted dengue virus which causes hemorrhagic fever (HF); encephalitis caused by Japanese encephalitis (JE) in tropical and subtropical region of the world. Around 50 - 100 million case of dengue fever are annually reported in more than 80 countries where the mosquito vector *Aedes aegypti* is endemic and approximately 500,000 patients suffer from dengue hemorrhagic fever and dengue shock syndrome. JE virus is the leading cause of arboviral encephalitis in Asia, accounting for 30,000 to 50,000 cases annually. St. Louis encephalitis virus causes sporadic epidemic encephalitis in the Americas. West Nile virus (WNV) has caused more than 9,000 cases in North America since 1999 **[1]**.

Flavivirus genome is a single-stranded, positive sense RNA of 10-11 kb containing a single ORF and is the only viral mRNA produced during the virus replication cycle. The replication takes place in the perinuclear region of cytoplasm in the infected cells. Three structural (capsid, premembrane and envelope) and seven nonstructural (NS1, NS2a, NS2b, NS3, NS4a, NS4b, and NS5) viral proteins are produced by proteolytic processing of the single polyprotein by viral and cellular proteases. The genome of *Flavivirus* open reading frame is flanked by the untranscribed region (UTR) in 5` and 3`. The UTR forms a complex RNA structure containing functional domain that are believed to play a role in virus translation, replication or assembly. These generates lots of scientific interest since, genetic modification within these region are known to attenuate *Flavivirus* without altering their antigenic specificity making them potential vaccine candidate for live attenuated vaccine **[2]**.

A wide range of algorithms has been developed to assist the identification of promoters in genomic sequence of many gene prediction methods. The regulatory element utilizing the TATT-box has been reported in the genome of Epstein - Barr virus (EBV). The motif was present in the promoters of lytic cycle genes and resembles a crucial host genome motif (TATA-box). Since the binding specificity of eukaryotic proteins recognizing TATA-box (TBP) was determined and no specific preference for interaction with TATT motif was found **[3]**. Consensus patterns of baculovirus sequences upstream from the translational initiation sites have been analyzed and a web tool. Potential consensus sequences such as TCATTGT, TCTTGTA, CTCGTAA, TCCATTT and TCATT plus TCGT in approximately 30 bp spacing context have been obtained **[4]**.

EIA and blCP0 are the activators for gene expression in adenovirus. It does not specifically bind both DNA and proteins interacting with chromatin remodeling enzymes. The capability of E1A to bind with Rb family members plays role in stimulating productive infection suggesting that E2F family members activate productive infection **[5].** Identification of transcription factor sites require for the function and regulation of genes. The evaluation of sequences specific DNA binding transcription factors belong to CTF/NF1 family has been reported **[6]**. Promoter is a most important regulatory region that controls and regulates gene expression at

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(4): 162-167 (2008)

the transcription level. It contains specific DNA sequences, response elements that are recognized by proteins known as transcription factors. These factors bind to the promoter sequences, recruiting RNA polymerase, the enzyme that synthesizes the RNA from coding region of the gene. However, a limited number of data is available on the promoter motifs in the genus *Flavivirus*. The present study was carried out to identify and analyze the putative promoter region present in *Flavivirus.*

Methodology:

Retrieval of genome sequence

The complete genome sequences of *Flavivirus* were retrieved from biological databases such as National Centre for Biotechnology Information cited at http://www.ncbi.nlm.nih.gov/genomes/VIRUSES/viruses.html and the Universal Virus Database of the International Committee on Taxonomy of Viruses (ICTVdB) genome database cited at http://www.ncbi.nlm.nih.gov/ICTVdb/.

BLAST

The relatedness of sequences deposited in databases was evaluated by BLAST (Basic Local Alignment Search Tool) implemented via the NCBI website (www.ncbi.nlm.nih.gov/blast) against the complete training dataset which is extracted from Genbank database. All these identified promoters were verified and searched for homology in the database.

Analysis of *Flavivirus* genome

The size of *Flavivirus* genome was analyzed with the aid of Generunner, DNAstar and ExPaSy tools. The G+ C content of each genome was also predicted (Table 1 under supplementary material).

Transcription start site

The transcription factor sites are over represented in the promoter region. It is natural to seek a prediction program based on putative TF site density. The PROMOTERSCAN program was used to identify the putative promoter in the genome of *Flavivirus*. This program comprises three database such as TF database, promoter database and non promoter set constructed from protein and RNA gene sequences. The density of all putative TF site is calculated separately for promoter and non promoter sequences scoring functions supplemented with a TATA matrix score **[7]**.

Results and discussion:

In the present study, complete genome sequences of 26 *Flavivirus* was analyzed. Within studied *Flavivirus*, the size of 24 genomes was approximately 10 kb. Merely, two genome sizes of Kamati River and Tick borne encephalitis virus was 11 kb. The smallest genome size is 10,053 bases of Tamana bat virus and the highest genome size (11,375 bases) is of Kamati river virus. The lower G+ C percentage of Tamana bat virus was 38.43 and the highest G+C content of Louping ill virus was 54.85. The genome size and G+C contents of *Flavivirus* were given (Table 1 under supplementary material). The establishment of persistent noncytopathic replication by replicon RNAs of a number of positive-strand RNA viruses usually leads to generation of adaptive mutations in nonstructural genes. Some of these adaptive mutations in hepatitis C virus increase the ability of RNA replication to resist the antiviral action of alpha/beta interferon (IFN-alpha/beta) and Sindbis virus may also lead to more efficient IFN production **[8]**.

Identification of important putative promoter in the complete genome of 26 medically significant *Flavivirus* was done. Total 22 different types of promoter were identified in the genomes and given (Table 2, see supplementary material). All these identified promoter motifs were confirmed with experimental obtained existing data. Sixteen viruses have the putative promoter sequences while ten *Flavivirus* do not have the promoter motifs in their genome. The identified putative promoter of *Flavivirus* was given (Table 3 under supplementary material). During the time of replication and multiplication all the host machinery was utilized by *Flavivirus* which did not have the promoter sequences.

The numbers of experimentally confirmed reports are available on the identification and characterization of promoter in the virus genome. The K1 gene of Kaposi's sarcoma-associated herpesvirus (KSHV) encodes a 46-kDa transmembrane glycoprotein that possesses transforming properties initiates signaling pathways in B cells and prevents apoptosis. K1 promoter demonstrated that purified Rta protein bound to the K1 at three locations independent of other DNA-binding factors **[9]**. Kaposi sarcoma-associated herpesvirus vIRF is a viral transcription factor that inhibits interferon signaling and transforms NIH 3T3 cells but does not bind interferon-stimulated response element (ISRE) DNA sequences **[10]**.

A eukaryotic promoter-specific activator protein (activators) stimulates the transcription. An acidic activator can directly interact with the transcription factor TFIIB and increase the stable assembly into a preinitiation complex **[11]**. Adult T-cell leukemia (ATL) is a complex and multifaceted disease associated with human T-cell leukemia virus type 1 (HTLV-I) infection. Viral oncoprotein is considered a major contributor to cell cycle deregulation in HTLV-I transformed cells by either directly disrupting cellular factors or altering their transcription profile. Tax transactivates these cellular promoters by interacting with transcription factors such as CREB/ATF, NF-kappaB, and SRF **[12]**. The transcription factor TFIID consisting of TATA-binding protein (TBP) and TBP-associated factors (TAFs) plays a central role in both positive and negative regulation of transcription. The TAF N-terminal domain (TAND) of TAF1 has been shown to interact with TBP and to modulate the interaction of TBP with the TATA box, which is required for transcriptional initiation and activation of TATA-promoter operated genes **[13]**.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(4): 162-167 (2008)

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

The regulatory element utilizing TATT box has been reported in the genome of Epstein - Barr virus (EBV). The motif is present in promoters of lytic cycle genes and resembled a crucial host genome motif (TATA-box). Since the binding specificity of eukaryotic proteins recognizing TATA-box (TBP) was determined and no specific preference for interaction with TATT motif was found **[3]**. Consensus patterns of baculovirus sequences upstream from the translational initiation sites have been analyzed and a web tool Local Alignment Promoter Predictor (LAPP) for the prediction of baculovirus promoter sequences has also been developed. Potential consensus sequences, i.e., TCATTGT, TCTTGTA, CTCGTAA, TCCATTT and TCATT plus TCGT in approximately 30 bp spacing context, have been found in baculovirus promoter regions, in addition to well characterized late and early promoter elements G/T/ATAAG and TATAA, which is accompanied about 30-bp downstream by a transcriptional initiation sequence CAGT or CATT **[4]**.

The adenovirus E1A gene and bICP0 encode proteins that are potent activators of viral gene expression. They do not specifically bind DNA and both proteins interact with chromatin-remodeling enzymes. A functional similarity of E1A was tested initially to see if it could stimulate BHV-1 productive infection. E1A consistently stimulates BHV-1 productive infection, but not as efficiently as bICP0. The ability of E1A to bind Rb family members plays a role in stimulating productive infection, suggesting that E2F family members activate productive infection. E2F-4, but not E2F-1, E2F-2 or E2F-5, activates productive infection with similar efficiency as E1A **[5]**.

Accurate prediction of transcription factor binding sites is needed to unravel the function and regulation of genes discovered in genome sequencing projects. To evaluate current computer prediction tools, we have begun a systematic study of the sequence-specific DNA-binding of a transcription factor belonging to the CTF/NFI family **[6]**. White Spot Syndrome Virus is a member of the virus family *Nimaviridae* and infecting shrimp and other crustacean species. The complete genome was analyzed to *in silico* identify the conserved promoter motifs. In the 5` upstream region contained the TATA box element is similar to the Drosophila RNA polymerase II core promoter sequences and utilization of the cellular transcription machinery for generating early transcripts **[10]**.

Conclusion:

The *in silico* identification of promoter motifs in the genome of *Flavivirus* was done. These promoters play vital role in the regulation of gene expression. Delineation of the promoter is fundamental for understanding gene expression patterns, regulation networks, cell specificity and development. It is also important for designing efficient expression vector or to target specific delivery system in the gene therapy. These results might help in designing the live attenuated vaccine candidate through the site directed mutagenesis in the promoter region. In the large scale genomic era promoter prediction is crucial for gene discovery and annotation.

References:

[01] J. L. Munoz-Jordan *et al.*, *J Virol*., 79: 8004 (2005)
[02] D. R. Durbin *et al. , Pediatrics.*, 108: E109 (2001)
[03] L. S. Wyrwicz *et al. , Antiviral Res.*, 75: 167 (2007)
[04] K. Xing *et al. , Virus Res.*, 113: 64 (2005)
[05] V. Geiser *et al. , J Gen Virol.*, 84: 929 (2003)
[06] E. Roulet *et al. , J Mol Biol.*, 297: 833 (2000)
[07] D.S. Prestridge, *J. Mol. Biol*., 249: 923 (1995)
[08] W. J. Liu *et al. , J. Virol.*, 78: 12225 (2004)
[09] B. S. Bowser *et al., Virol.*, 348: 309 (2006)
[10] S. Jayachandra *et al. , Proc Natl Acad Sci*., 96: 11566 (1999)
[11] S. G. Roberts *et al. , Green Nature*, 363: 741 (1993)
[12] C. de la Fuente *et al. , Retrovirol*., 5: 43 (2006)
[13] T. K. Mal *et al. , J Mol Biol*., 339: 681 (2004)

Edited by P. Kangueane

Citation: Somvanshi *et al.*, Bioinformation 3(4): 162-167 (2008)

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(4): 162-167 (2008)

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Supplementary material

Designation of virus	Accession No.	Size (bp)	G+C (%)
Dengue virus 1	AB178040	10,735	46.55
Dengue virus 2	NC_001474	10,723	45.82
Dengue virus 3	NC_001475	10,707	46.71
Dengue virus 4	NC_002640	10,649	47.12
Japanese encephalitis virus	NC_001437	10,976	51.42
Kokobera virus	NC_009029	10,233	49.62
Modoc virus	NC_003635	10,600	45.56
Ilheus virus	AY632539	10,275	52.29
Sepik virus	NC_008719	10,793	47.29
Yokose virus	NC_005039	10,857	47.01
Apoi virus	NC_003676	10,116	48.38
Entebbe bat virus	NC_008718	10,510	50.47
Alkhurma virus	NC_004355	10,685	54.06
Langat virus	NC_003690	10,943	54.31
Louping ill virus	NC_001809	10,871	54.85
Omsk hemorrhagic fever virus	NC_005062	10,787	53.64
Powassan virus	NC_003687	10,839	53.32
Karshi virus	NC_006947	10,653	54.41
Tick-borne encephalitis virus	NC_001672	11,141	53.77
Yellow fever virus	NC_002031	10,862	49.72
Bussuquara virus	NC_009026	10,290	49.85
Cell fusing agent virus	NC_001564	10,695	51.31
Iguape virus	NC_009027	10,251	50.73
Montana myotis leukoencephalitis virus	NC_004119	10,690	44.15
Kamiti River virus	NC_005064	11,375	50.29
Tamana bat virus	NC_003996	10,053	38.43

Table 1: Diverse medically important *Flavivirus* used in this study

Promoter designation	Promoter sequences	Bases
INF.1	AAGTGA	6
beta-pol CS	NTGACGTCAN	10
ATF RS	TGACGT	6
CRE.1	CGTCA	5
CREB CS1	ACGTCA	6
UCE.2	GGCCG	5
CTF/NF-1a	AGCCAAT	7
Sp1-IE-3.4	CCGCCC	6
Sp1-SV40.4	GGGCGG	6
AP-2 CS6	CCCMNSSS	8
Sp1-IE-3.3	CCCGCC	6
Sp1-hsp70 (1)	GGCGGG	6
T-Ag-SV40.3	GGGGC	6
E2F-myc	GCGGGAAAA	9
E2F CS	TTTTSSCGS	9
USF-IIEII	TGACGCA	7
ATF CS2	TGACGYMR	8
TFIID (TATA-box.2)	TATAAA	6
NF-kB-hIL2Ra	GGGGAATCTCC	11

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(4): 162-167 (2008)

MBF-I CS	TGCRCRC	7
CTF/NF-1a	AGCCAAT	7
NFkB CS4	GGGRNTYYC	9

Table2: The promoter sequences identified in this study.

Virus designation	Promoter designation	Specific location	Ranged of location in genome	Strand
Dengue virus 1	MBF-I	5169	5206 to 4956	Reverse
	UCE.1	5144		
	T- Ag	5045		
	ATF	5006		
Dengue virus 2	AP-2	5473	5381 to 5631	Forward
	Sp1	5476		
	CTF	5517		
	TFIID	5613		
Dengue virus 3	**ND**			
Dengue virus 4	INF.1	4507	4499 to 4749	Forward
	CTF	4714		
	beta-pol_CS	4732		
	ATF/CREB	4738		
Japanese encephalitis	UCE.2	4382	4355 to 4605	Forward
	CREB	4461		
	CTF/NF-1	4533		
	Sp1	4537		
Kokobera virus	**ND**			
Modoc virus	Sp1	3502	3261 to 3511	Forward
	E2F	3511		
Ilheus virus	EIIF	9540	9773 to 9523	Reverse
	E2F	9533		
Sepik virus	**ND**			
Yokose virus	**ND**			
Apoi virus	**ND**			
Entebbe bat virus	T-Ag	1020	893 to 1143	Forward
	AP-2	1024		
	USF-IIEII	1074		
	TFIID	1127		
Alkhurma virus	Sp1	4226	4185 to 4435	Forward
	T-Ag	4294		
	CBF	4404		
	INF.1	4415		
	NF-kB	4434		
Langat virus	T-Ag	1201	1186 to 1436	Forward
	MBF-I	1247		
	AP-2	1274		
	USF-IIEII	1322		
	ATF	1322		
	CTF/NF-1	1426		
Louping ill virus	AP-2	10503	10425 to 10675	Forward
	Sp1	10640		
	UCE.2	10657		
Omsk hemorrhagic fever virus	**ND**			

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(4): 162-167 (2008)

Powassan virus	**ND**			
Karshi virus	AP-1	6817	6897 to 6647	Reverse
	CTF/NF-1	6795		
	CTF	6773		
	NFI	6752		
	Sp1	6668		
Tick-borne encephalitis virus	**ND**			
Yellow fever virus	CTF	10198	10196 to 10446	Forward
	AP-1	10247		
	TFIID	10430		
Bussuquara virus	UCE.2	10179	10287 to 10037	Reverse
	CTF	10134		
	GATA-1	10083		
Cell fusing agent virus	element_II_rs-4	840	840 to 1090	Forward
	T-Ag	1088		
	T-Ag	6230	6037 to 6287	Forward
	AP-2	6231		
	CTF	6264		
Iguape virus	CTF/NF-1	4422	4416 to 4666	Forward
	T-Ag	4477		
	Sp1	4538		
	ATF/CREB	4625		
	Sp1	4640		
	GCF	4657		
	UCE.2	4660		
Montana myotis leukoencephalitis virus	**ND**			
Kamiti River virus	INF.1	79	20 to 270	Forward
	T-Ag	120		
	Sp1	160		
	E2F	172		
	USF-IIEII	10982	10982 to 11232	Forward
	CREB	10986		
	AP-2	11143		
	TFIID	11205		
Tamana bat virus	ND			

Table 3: Identification of putative promoter region in *Flavivirus* genome. **ND**: Not detected promoter sequences in the *Flavivirus*.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(4): 162-167 (2008)