RESEARCH NOTE

Abundance, composition and distribution of simple sequence repeats and dinucleotide compositional bias within WSSV genomes

MALATHI SHEKAR, INDRANI KARUNASAGAR and IDDYA KARUNASAGAR*

Department of Fishery Microbiology, UNESCO Centre for Marine Biotechnology, Karnataka Veterinary, Animal and Fishery Sciences University, College of Fisheries, Mangalore 575 002, India

Introduction

White spot syndrome virus (WSSV), recognized by the International Committee on Taxonomy of Viruses (ICTV) as a member of a new family Nimaviridae and genus *Whispovirus*, is one of the largest animal viruses, with a genome size of ~300 kilobases, and is responsible for serious losses of cultured shrimp worldwide. Three complete genome sequences of the virus isolated from Thailand (Van Hulten *et al*. 2001), China (Yang *et al*. 2001) and Taiwan (Tsai *et al*. 2000), are available in the GenBank database, but have been used for very few comparative studies. A high degree of genetic similarity between the isolates has been demonstrated by an RFLP study (Wang *et al*. 2000). Marks *et al*. (2004) showed genetic variation among the three isolates of WSSV and reported deletions of ~13 kb and 1 kb in the Thailand and China isolates, respectively, in comparison to the Taiwan isolate.

This study aims to study genetic differences in the three WSSV isolates using simple sequence repeats (SSRs), or microsatellites, which are DNA motifs of 1–6 bases tandemly repeated several times. SSRs are ubiquitous, being found in varying abundance in both eukaryotic (Gur-Arie *et al*. 2000) and prokaryotic (van Belkum *et al*. 1998) genomes, and are considered hotspots of mutation. Differential patterns of SSRs in coding and noncoding regions of DNA reflect a difference in evolutionary pressure on various functional parts of the genome (van Belkum *et al*. 1998). Variable SSRs within genes or regulatory regions may influence gene expression, providing an evolutionary advantage through fast adaptation to changing environmental conditions. Against this background, we have analysed the three WSSV genome sequences and done a comparative analysis with respect to the abundance, distribution and composition of SSRs which could be informative in the study of functional effects, evolutionary relationships, and adaptation to different hosts. We also report the compositional bias of the three WSSV genomes and the genetic distances among them.

*For correspondence. E-mail: mircen@sancharnet.in, karuna8sagar@yahoo.com.

Materials and methods

The three complete WSSV genome sequences (accession numbers AF369029, AF332093, AF440570) were downloaded from GenBank. The sizes of these genomes are 292,967 bp, 305,107 bp and 307,287 bp and they have been designated W-29, W-93 and W-70 for the Thailand, China and Taiwan isolates, respectively. Mononucleotide SSRs in the genomes were screened using the *motif search* program of Sequence Quickie-Calc™ version 5.0 software, whereas for dinucleotide, trinucleotide, tetranucleotide, pentanucleotide and hexanucleotide SSRs, we used the software SSRIT (http://www.gramene.org/db/searches/ssrtool). The minimal repeat number was taken as three.

Statistical analysis of SSR frequencies was done by constructing 10 randomized simulated genomes for each WSSV sequence, using the *shuffleseq* program of EMBOSS package (http://sequenzanalyse.biologie.uni-konstanz.de/EMBOSS/index.html). Each simulated genome was then analysed for SSRs of a given composition and size using the software mentioned above. Deviation of observed numbers of SSRs of given motif and repeat number from expectations were tested using two-tailed *t*-tests. To determine the distribution of SSRs among coding and noncoding regions, the coding regions were extracted and parsed into a new file and SSRs determined.

Keywords. shrimp; white spot syndrome virus (WSSV); simple sequence repeats (SSRs); compositional bias; genetic distance.

Compositional bias among dinucleotides was determined for all WSSV genomes based on the calculation of Karlin *et al*. (1997). The whole genome sequence was concatenated with its inverted complementary sequence using the procedures *revseq* and *union* (EMBOSS package). Mononucleotide and dinucleotide frequencies were calculated using the *compseq* procedure (EMBOSS package). Dinucleotide relative abundance profile p^*_{XY} was calculated using the equation $p^*_{XY} = f^*_{XY}/f^*_X f^*_Y$, where f_{XY} is the frequency of dinucleotide XY, and f_X and f_Y the frequencies of nucleotides X and Y. Statistical analysis from previous studies indicates the normal expected values of p^*_{XY} to be in the range of 0.78–1.23 for double-stranded DNA. Degress of dinucleotide over-representation are grouped into $+ (1.23 \leq p* < 1.30)$, $++$ $(1.30 \leq p* < 1.50)$ and $+++$ $(p* \geq 1.50)$, while underrepresentation is grouped into $-$ $(0.70 < p* < 0.78)$, $--$ $(0.50 < p* \leq 0.70)$ and $---$ $(p* \leq 0.50)$. A measure of the difference between two whole genome sequences was calculated using the equation

$$\delta^*(f, g) = 1/16 \sum \left| {p^*_{XY}}^{(f)} - {p^*_{XY}}^{(g)} \right| \times 1000,$$

where δ^* is the dinucleotide relative abundance distance between two sequences f and g, X and Y are the nucleotides A, T, C or G, and the sum extends over all dinucleotides. This is the average absolute dinucleotide relative distance, referred to as δ^*-distance, and was calculated by dividing the whole genome sequence into six nonoverlapping 50-kb fragments and calculating p^*_{XY} for each fragment. The average $\delta^*(f, g)$ between fragments was calculated. Two-sample t-tests were performed to compare mean δ^* distances of 50-kb regions.

Results and discussion

Our study revealed a large number of SSRs scattered throughout the entire WSSV genomes. The total number of nucleotides encompassed in SSRs was 70,364, 73,503 and 73,879 bases for the Thailand, China and Taiwan isolates, respectively, which comprised ~24% of the entire genome. On examining the SSR unit size classes, mononucleotide repeats were the most abundant (20.4%), followed by trinucleotide (2.1%) and dinucleotide repeats (1.4%) in all the genomes. Tetranucleotide and hexanucleotide repeats were least in number and did not repeat more than four times. There were no SSRs with pentanucleotide repeats (table 1).

From table 1, it can be seen that in the three genomes, mononucleotide repeats were repeated a maximum of 11 times and dinucleotides a maximum of six times. Trinucleotides, on the other hand, had variable representation. In W-29 and W-93 genomes, repeat tracts of 11 were recorded, whereas in W-70 the repeat number was 12. The minor difference was due to polymorphism observed in one of the repeats. Length distribution of all SSRs indicated that the frequency of repeats decreased exponentially with repeat length. This is probably because longer repeats are known to be highly unstable and have a higher mutation rate (Kruglyak *et al*. 1998). The repeat distribution showed an excess of shorter repeats, up to four times among trinucleotides and seven times among mononucleotides, compared to randomized genomes of identical nucleotide composition and size ($P < 0.001$).

Repeats were distributed in both coding and noncoding regions in all three WSSV genomes. It was observed that as mononucleotide repeat length increased, such repeats were found almost exclusively in the noncoding regions. Dinucleotide repeats were present in both coding and noncoding regions, while trinucleotides were observed to be overrepresented in the coding regions (table 2).

A cursory glance at the nucleotide compositions of SSR tracts shows that poly(G) and poly(C) repeats are underrepresented in coding and noncoding regions in the three genomes, while poly(A) is overrepresented in coding regions and poly(T) in noncoding regions (see table in electronic supplementary material at http://www.ias.ac.in/jgenet). This assumes significance as pure A and T tracts are associated with decreased base-pair dissociation constants, thereby increasing DNA stability, while increased length for G tracts promotes base-pair-dissociation (Dornberger *et al*. 1999).

Dinucleotide repeats are estimated to have the highest slippage rates among SSRs (Chakraborty *et al*. 1997). Of the six dinucleotide combinations, we observed the frequency of AG/GA repeat motif to be highest (28.5–30.5%) in coding regions, while AT/TA (25.5–27.8%) was most abundant in noncoding regions, followed by AC/CA (20.3–23.5%). Expectations of dinucleotide frequencies were calculated on the basis of genomewide nucleotide composition described by Gur-Arie *et al*. (2000) and all values clearly exceeded the expected value. Bachtrog *et al*. (1999) detected a significant correlation between AT content and AT/TA SSR repeats, suggesting that the SSR genesis may be a random process. This is of particular importance as strand separation is less likely at these AT repeat regions due to their lower stacking interactions, compared to GC tracts, thus increasing slipped strand mispairing. Dinucleotide arrangements within genomes are known to influence many structural properties of DNA which play an important role in biological process like replication, such as curvature flexibility and helix stability. However, the exact function of AG repeats in coding regions and of AC repeats in noncoding regions needs to be investigated in WSSV genomes.

Of the 64 triplet repeat types, the density of GAA, TCT and AAC was the highest (see figure 1 in electronic supplementary material at http://www.ias.ac.in/jgenet). In protein sequences, a single motif repeated tandemly can form a part of a secondary structure, giving rise to conformations such as an α/α coil, or right-handed and left-handed β-helices (Herringa and Taylor 1997). Triplets coding for serine were found to be the most abundant, followed by triplets for glutamic acid and leucine (see figure 2 in electronic supplementary material at http://www.ias.ac.in/jgenet). Glutamic acid motifs were predominant in the coding regions of the three

Table 1. Comparative distribution of SSRs within the three WSSV genomes (W-29, Thailand; W-93, China; W-70, Taiwan).

Repeats	Nucleotide repetition unit (motif)																	
per	mono			di			tri			tetra			penta			hexa		
locus	W-29	W-93	W-70	W-29	W-93	W-70	W-29	W-93	W-70	W-29	W-93	W-70	W-29	W-93	W-70	W-29	W-93	W-70
3	11,379^{+}	11,956^{+}	11,817^{+}	629	653	657	402^{+}	413^{+}	417^{+}	5	5	5	–	–	–	10	11	11
4	3790^{+}	3910^{+}	4037^{+}	39	39^{-}	40^{-}	95^{+}	100^{+}	101^{+}	1	1	1	–	–	–	1	1	1
5	1471^{+}	1524^{+}	1554^{+}	1	4	4	47	49	48	–	–	–	–	–	–	–	–	–
6	377^{+}	407^{+}	427^{+}	1	1	1	23	23	24	–	–	–	–	–	–	–	–	–
7	79^{+}	82^{+}	80^{+}	–	–	–	8	10	8	–	–	–	–	–	–	–	–	–
8	25^{+}	27	22	–	–	–	6	5	7	–	–	–	–	–	–	–	–	–
9	4	4	7	–	–	–	–	1	1	–	–	–	–	–	–	–	–	–
10	1	3^{+}	1	–	–	–	1	1	–	–	–	–	–	–	–	–	–	–
11	1	1	1	–	–	–	1	1	–	–	–	–	–	–	–	–	–	–
12	–	–	–	–	–	–	–	–	1	–	–	–	–	–	–	–	–	–

$^{+}$significantly more, and $^{-}$ significantly less than mean frequencies in computer-generated randomized genomes at $P < 0.0001$.

Table 2. Distribution of SSRs among coding and noncoding regions of the three WSSV genomes.

	Thailand (W-29)					China (W-93)					Taiwan (W-70)				
	Total	Coding		Noncoding		Total	Coding		Noncoding		Total	Coding		Noncoding	
	No.	No.	%	No.	%	No.	No.	%	No.	%	No.	No.	%	No.	%
Mononucleotides															
3	11,379	5957	52.4	5422	47.6	11,956	7010	59.7	4737	40.3	11,817	7424	62.8	4393	37.2
4	3790	2028	53.5	1762	46.5	3910	2451	61.5	1532	38.5	4037	2542	63.0	1495	37.0
5	1471	699	47.5	772	52.5	1524	856	55.9	674	44.1	1554	904	58.2	650	41.8
6	377	169	44.8	208	55.2	407	207	52.1	190	47.9	427	213	49.9	214	50.1
7	79	41	51.9	38	48.1	82	51	60.7	33	39.3	80	49	61.3	31	38.7
8	25	7	28.0	18	72.0	27	15	51.7	14	48.3	22	8	36.4	14	63.6
9	5	–	–	5	100.0	4	–	–	4	100.0	7	1	14.3	6	85.7
10	1	–	–	1	100.0	3	1	25.0	3	75.0	1	–	–	1	100.0
11	1	–	–	1	100.0	1	–	–	1	100.0	1	–	–	1	100.0
Dinucleotides ≥6 bp	670	325	48.5	345	51.5	697	386	55.4	311	44.6	702	411	58.5	291	41.5
Trinucleotides ≥6 bp	583	332	57.0	251	43.0	603	409	67.8	194	32.2	607	414	68.2	193	31.8
Tetranucleotides	5	1	20.0	4	80.0	6	3	50.0	3	50.0	6	2	33.3	4	66.7
Pentanucleotides	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Hexanucleotides	10	5	50.0	6	50.0	12	10	83.3	2	16.7	12	10	83.3	2	16.7
Genome partition			53.4		46.6			62.4		37.6			63.5		36.5

genomes, whereas there was a discrepancy in the distribution of serine and leucine motifs between coding and noncoding regions. Serine class repeats were predominant in the coding regions of W-93 and W-70, whereas in W-29 genome serine repeats were abundant in the noncoding regions. In the case of leucine repeats, in W-29 they were more abundant in noncoding regions, but were nearly equally distributed among coding and noncoding regions in W-93 and W-70. It is possible that certain mutational processes leading to expansion and contraction of nucleotide repeats enable certain triplet tracts to function in different ways.

We also observed that WSSV genomes have very short repeat tracts, contrary to what is seen in bacteria. Such observations have also been made in human cytomegalovirus genome, and such tracts can serve as polymorphic markers (Davis *et al*. 1999). Polymorphism existing within WSSV trinucleotide repeats has been reported earlier (Shekar *et al*. 2005), raising the possibility of using them as molecular markers. Shorter repeats in viral genomes could be involved in events such as recombination, replication and repair mechanisms, possibly helping the organism to adapt to different hosts. Except for minor differences, the three genomes were largely similar in the distribution and composition of SSRs.

Genomewide dinucleotide compositional bias comparison has been used as a tool to study evolution (Karlin *et al*. 1997). Dinucleotide relative abundance profiles of related organisms are similar, whereas they tend to be different in distantly related organisms (Karlin *et al*. 1997). Within genomes, the abundance values are relatively constant in both coding and noncoding regions, giving a genomewide perspective of the nucleotide composition patterns, making it possible to compute a 'genome signature' for several organisms (Blaisdell *et al*. 1996). Genomic signatures are useful in phylogenetic analyses (Karlin *et al*. 1998), genetic sequence classification (Chuzhanova *et al*. 1998), and classification of genes from different genomes (Nakashima *et al*. 1998). Moreover, dinucleotide biases within a genome might influence DNA replication and repair machinery, contributing significantly to the generation and maintenance of species-specific dinucleotide relative abundance (Karlin *et al*. 1997). Dinucleotide abundance values in the three WSSV genomes are shown in table 3. The dinucleotides TA (p^*_{TA} = 0.721) and CG (p^*_{CG} = 0.640) were moderately underrepresented in the three genomes, while all others were in the normal range. It has been suggested that TA suppression relates to low tyrosine usage and minimizes inappropriate binding of transcription or termination factors (Burge *et al*. 1992), whereas CG suppression relates to high mutability of methylated cytosine in vertebrate sequences (Bird 1980). A study by Karlin *et al*. (1994) showed CG to be underrepresented only in small viral genomes and not in large eukaryotic viruses. On the contrary, we found CG underrepresentation in our study of WSSV, an animal virus with a large genome of ~300 kb.

A comparison of all the three WSSV genomes based on the dinucleotide relative abundances revealed the genetic distance between the three isolates to be very small (δ^*(W-29, W-93) = 1.0, δ^*(W-29, W-70) = 1.25; δ^*(W-93, W-70 = 0.75)) even though they originate from different geographical regions. We can, therefore, infer that the WSSV sequences are closely related by ancestry.

Table 3. Dinucleotide relative abundance in the three WSSV genomes (W-29, Thailand; W-93, China; W-70, Taiwan).

	W-29		W-93		W-70	
XY	p^*_{XY}	Over-under-represented	p^*_{XY}	Over-under-represented	p^*_{XY}	Over-under-represented
AA	1.115		1.115		1.117	
AC	0.905		0.905		0.905	
AG	1.058		1.056		1.057	
AT	0.911		0.912		0.910	
CA	1.120		1.119		1.118	
CC	1.108		1.109		1.109	
CG	0.638	– –	0.639	– –	0.640	– –
CT	1.058		1.056		1.057	
GA	1.115		1.116		1.116	
GC	0.864		0.860		0.861	
GG	1.108		1.109		1.109	
GT	0.905		0.905		0.905	
TA	0.721	–	0.721	–	0.721	–
TC	1.115		1.116		1.116	
TG	1.120		1.119		1.118	
TT	1.115		1.115		1.117	

*Significant underrepresentation shown as '–' ($0.70 < p^* < 0.78$) and '– –' ($0.50 < p^* \leq 0.70$). All the other values are within the normal range.

Furthermore, based on the δ^* values obtained between the three isolates we can conclude that WSSV sequences from China and Taiwan have the closest relationship among the three isolates.

Acknowledgements

Financial support from the Department of Biotechnology, Government of India, under the Bioinformatics Centre programme is gratefully acknowledged.

References

Bachtrog D., Weiss S., Zangerl B., Brem G. and Schlötterer C. 1999 Distribution of dinucleotide microsatellites in the *Drosophila melanogaster* genome. *Mol. Biol. Evol*. **16**, 602–610.

Bird A. P. 1980 DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res*. **8**, 1499–1504.

Blaisdell B. E., Campbell A. M. and Karlin S. 1996 Similarities and dissimilarities of phage genomes. *Proc. Natl. Acad. Sci. USA* **93**, 5854–5859.

Burge C., Campbell A. M. and Karlin S. 1992 Over- and underrepresentation of short oligonucleotides in DNA sequences. *Proc. Natl. Acad. Sci. USA* **89**, 1358–1362.

Chakraborty R., Kimmel M., Stivers D. N., Davidson L. J. and Deka R. 1997 Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. *Proc. Natl. Acad. Sci. USA* **94**, 1041–1046.

Chuzhanova N. A., Jones A. J. and Margett S. 1998 Feature selection for genetic sequence classification. *Bioinformatics* **14**, 139–143.

Davis C. L., Field D., Metzgar D., Saiz R., Morin P. A., Smith I. L., Spector S. A. and Wills C. 1999 Numerous length polymorphisms at short tandem repeats in human cytomegalovirus. *J. Virol*. **73**, 6265–6270.

Dornberger U., Leijon M. and Fritzsche H. 1999 High base pair opening rates in tracts of GC base pairs. *J. Biol. Chem*. **274**, 6957–6962.

Gur-Arie R., Cohen C. J., Eitan Y., Shelef L., Hallerman E. M. and Kashi Y. 2000 Simple sequence repeats in *Escherichia coli*: abundance, distribution, composition and polymorphism. *Genet. Res*. **10**, 62–71.

Herringa J. and Taylor W. R. 1997 Three-dimensional domain duplication, swapping and stealing. *Curr. Opin. Struct. Biol*. **7**, 416–421.

Karlin S., Doerfler W. and Cardon L. R. 1994 Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? *J. Virol*. **68**, 2889–2897.

Karlin S., Mrazek J. and Campbell A. M. 1997 Compositional biases of bacterial genomes and evolutionary implications. *J. Bacteriol*. **179**, 3899–3913.

Karlin S., Campbell A. M. and Mrazek J. 1998 Comparative DNA analysis across diverse genomes. *Annu. Rev. Genet*. **32**, 185–225.

Kruglyak S., Durrett R. T., Schug M. D. and Aquadro C. F. 1998 Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc. Natl. Acad. Sci. USA* **95**, 10774–10778.

Marks H., Goldbach R. W., Vlak J. M. and van Hulten M. C. W. 2004 Genetic variation among isolates of white spot syndrome virus. *Arch. Virol*. **149**, 673–697.

Nakashima H., Ota M., Nishikawa K. and Ooi T. 1998 Genes from nine genomes are separated into their organisms in the dinucleotide composition space. *DNA Res*. **5**, 251–259.

Shekar M., Karunasagar I. and Karunasagar I. 2005 A computer based identification of variable number of tandem repeats in white spot syndrome virus genomes. *Curr. Sci*. **89**, 882–887.

Tsai M. F., Lo C. F., van Hulten M. C., Tzeng H. F., Chou C. M., Huang C. J. *et al*. 2000 Transcriptional analysis of the ribonucleotide reductase genes of shrimp white spot syndrome virus. *Virology* **277**, 92–99.

van Belkum A., Scherer S., van Alphen L. and Verbrugh H. 1998 Short-sequence DNA repeats in prokaryotic genomes. *Microbiol. Mol. Biol. Rev*. **62**, 275–293.

van Hulten M. C., Witteveldt J., Peters S., Kloosterboer N., Tarchini R., Fiers M. *et al*. 2001 The white spot syndrome virus DNA genome sequence. *Virology* **286**, 7–22.

Wang Q., Nunan L. M. and Lightner D. V. 2000 Identification of genomic variations among geographic isolates of white spot syndrome virus using restriction analysis and Southern blot hybridization. *Dis. Aquat. Org*. **43**, 175–181.

Yang F., He J., Lin X., Li Q., Pan D., Zhang X. and Xu X. 2001 Complete genome sequence of the shrimp white spot bacilliform virus. *J. Virol*. **75**, 11811–11820.

Received 24 February 2006