Codon adaptation index analysis of RNA genome plant viruses

The codon adaptation index (CAI) was proposed as a quantitative way of predicting the expression level of a gene based on its codon sequence[1]. Expression level indicators such as CAI are widely used and are important in a variety of contexts. First, these indicators can serve as one of the variables to determine how likely is the transcription and translation of an open reading frame (ORF) into a protein product. Secondly, in heterologous gene expression, codon-based expression indicators are helpful in finding codon sequences that are most likely to yield high expression.

The CAI model assigns a parameter, termed 'relative adapativeness' to each of the 61 codons (stop codons excluded)[1]. The relative adaptiveness of codon is defined as its frequency relative to the most often used synonymous codon; note that this parameter is computed from a set of highly expressed genes G. It is given by:

$$w_{\mathrm{aa},i}(G) = \frac{f_{\mathrm{aa},i}(G)}{f_{\mathrm{aa,max}}(G)}, \qquad (1)$$

where $f_{\mathrm{aa},i}$ is the frequency of codon i (which encodes amino acid, aa) and $f_{\mathrm{aa,max}}$ the frequency of the codon most used for encoding amino acid aa in a set of highly expressed genes G. The relative adaptiveness parameter $w_{\mathrm{aa},i}$ ranges from 0 to 1, with 0 indicating that codon is not present at all in G, and 1 indicating a codon that occurs most often in G for a given amino acid.

The CAI of gene g is then simply the geometric average of the relative adaptiveness of all the codons in a gene sequence:

$$\mathrm{CAI}_g = \prod_{i=1}^{N} w_i^{1/N}, \qquad (2)$$

where w_i is relative adaptiveness of the ith codon in a gene with N codons. This formula can be transformed into:

$$\mathrm{CAI}_g = \prod_{k=1}^{61} w_k^{X_{k,g}}, \qquad (3)$$

where w_k now represents the relative adaptiveness of the kth codon among 61 codons in the genetic code and $X_{k,g}$ is the fraction of codon k among the total number of codons in gene g:

$$X_{k,g} = \frac{C_{k,g}}{\sum_{i=1}^{61} C_{i,g}}, \qquad (4)$$

where $C_{k,g}$ is the number of times codon k appears in gene g. Note that $w_k = w_k(G)$ in eq. (3) is dependent on the set of highly expressed genes G. Like relative adaptiveness, CAI also ranges from 0 to 1. Higher CAI values indicate genes that are more likely to be highly expressed.

The genome composition of living organisms can vary widely. This is considered to be the result of the directional mutational bias towards GC or AT[2]. This bias could theoretically be due to a bias in the copying error of viral RNA polymerase, selection pressure, or editing by host RNA-editing enzymes. Certain types of hyperpermutation have been described in a number of viruses[3], and may also contribute to viral genome composition.

The GC content of a genome has been shown to be a major contributing factor to the codon usage bias, which could affect expression efficiency[4–7]. It is interesting to see how GC content interacts with genome polarity and codon usage bias in RNA viruses. Genome composition and codon usage bias are particularly interesting in the RNA viruses because the same RNA may be used as mRNA, genome or antigenome. Replication of the RNA genome is also different from DNA replication of the host using different polymerase enzymes and in different environments, which may contribute to mutational bias that drives the genome composition. RNA viruses with positive and negative-stranded genome are different in their strategies of genome expression and replication, which may contribute to mutational bias and selection pressure.

For analyses, we retrieved the genomic sequences and coding sequences of 73 plant viruses from the NCBI database. To calculate the GC content we used the software tool BIOEDIT. CAI was calculated on the server of the Evolving Code Group at the University of Maryland, USA (http://www.evolvingcode.net/codon/CAI_Calculator.php).

The RNA viruses were chosen to cover most viral families/groups causing diseases of economic importance in plants. Names of viruses and their genome compositions are given in Table 1. There is significant difference in the GC content of positive-stranded RNA vs negative-stranded RNA viruses. The positive-stranded viruses have a mean GC content of 45.12%, while that of negative-stranded RNA viruses is 35.79%. The double-stranded RNA viruses have GC content of 42.10%. The highest GC content of 66.23% was found in Grapevine fleck virus, which is a monopartite positive stranded virus, whereas lowest GC content of 31.91% was found in Fiji disease virus, which has a double-stranded RNA genome. In general, monopartite (GC – 47.59%) positive-stranded RNA viruses have higher GC content over bipartite (GC – 43.45%) and tripartite (GC – 44.33%).

To study the codon bias in relation to predicted transitional efficiency in plant cells, we calculated CAI values using highly expressed host genes as the reference set[8]. This highly expressed codon set has been used successfully for codon optimization in viral genes.

CAIs varied widely among viruses ranging from 0.44 (in Clover yellow mosaic virus) to 0.823 (in Maize rayado fino virus) for positive-stranded RNA viruses, and 0.342 (in Impatiens necrotic spot virus) to 0.512 (in Lettuce ring necrosis virus) for negative-stranded viruses. The average CAI value for positive-stranded RNA viruses was 0.666, while that for negative-stranded viruses was found to be 0.406. This confirmed that mainly GC content drives codon bias of RNA viruses and consequently, the positive-stranded RNA viruses had higher CAI value than the negative-stranded viruses.

In this set of RNA viruses, GC content correlated with the CAIs value, with a Pearson correlation coefficient of 0.959 ($P < 0.01$). This result confirmed that codon bias of RNA viruses is driven mainly by GC content, and consequently the positive-stranded viruses have higher CAI than the negative-stranded viruses (0.823 versus 0.342, $P < 0.001$, t-test). Since codons contain different GC content, the amino acid content can be biased by the GC content. To determine

Table 1. Codon adaptation index and GC content of plant RNA viruses

Virus	Size (N)	CAI	GC%	GARP%	A%	C%	G%	T%
Tombusviridae								
Pothos latent virus (I)	4354	0.538	47.34	26.34	25.70	21.34	26.00	26.96
Oat chlorotic stunt virus (I)	4114	0.538	50.49	26.08	24.21	24.33	26.15	25.30
Carnation mottle virus (I)	4003	0.781	48.79	26.27	28.33	21.53	27.25	22.88
Red clover necrotic mosaic virus (II)	5338	NC	46.57	NC	28.92	22.12	24.45	24.50
Maize chlorotic mottle virus (I)	4437	0.698	50.19	27.74	27.00	25.13	25.06	22.81
Panicum mosaic virus (I)	4327	NC	50.02	NC	28.11	25.73	24.29	21.87
Cucumber necrosis virus (I)	4701	0.704	48.88	24.15	26.12	21.42	27.46	24.95
Tomato bushy stunt virus (I)	4776	NC	48.12	NC	26.32	20.58	27.53	25.54
Melon necrotic spot virus (I)	4262	0.714	45.89	22.76	24.80	20.95	24.94	29.28
Carnation Italian ring spot virus (I)	4760	0.800	48.11	27.15	26.41	20.78	27.73	25.48
Tymoviridae								
Grapevine fleck virus (I)	7564	0.780	66.23	35.27	13.93	49.89	16.34	19.83
Maize rayado fino virus (I)	6305	0.823	61.98	34.13	15.29	38.37	23.62	22.73
Turnip yellow mosaic virus (I)	6318	NC	56.44	NC	22.84	39.38	17.06	20.28
Bromoviridae								
Brome mosaic virus (II)	6099	0.518	46.06	22.44	26.25	21.04	25.02	27.69
Cowpea chlorotic mottle virus (III)	8118	0.767	43.52	21.78	27.65	19.44	24.08	28.82
Cowpea mottle virus (I)	4029	0.707	51.35	27.27	25.34	25.34	26.01	23.28
Cucumber mosaic virus (III)	8863	0.728	47.08	27.74	24.45	23.08	24.00	28.47
Tobacco streak virus (III)	8622	0.718	43.38	21.75	27.88	20.40	22.98	28.74
Olive latent virus-2 (III)	8301	0.471	48.20	25.83	24.47	21.97	26.33	27.33
Alfalfa mosaic virus (III)	8274	0.677	42.69	23.02	27.96	20.75	21.94	29.36
Caulimoviridae								
Carnation etched ring virus (I)	7932	0.816	36.36	16.96	37.03	18.18	18.18	26.61
Closteroviridae								
Beet yellows virus (I)	15480	NC	46.03	NC	25.14	22.26	23.77	28.83
Grapevine leafroll virus-3 (I)	17919	NC	46.14	NC	26.39	19.63	26.45	27.47
Lettuce infectious yellows virus (I)	8118	0.669	36.62	16.39	34.58	15.96	20.66	28.80
Flexiviridae								
Apple chlorotic leaf spot virus (I)	7545	0.635	42.13	21.20	31.48	17.92	24.21	26.39
Apple stem grooving virus (I)	6495	0.604	41.45	20.95	30.56	18.41	23.03	27.99
Grapevine virus-A (I)	7349	0.801	49.04	22.01	29.87	21.64	27.44	21.09
Indian citrus ring spot virus (I)	7560	0.615	51.96	26.05	27.96	32.33	19.63	20.08
Pear black necrotic leaf spot virus (I)	6497	NC	42.31	NC	30.35	18.76	23.55	27.34
Rupestris stem pitting associated virus (I)	8744	NC	42.92	NC	27.79	19.19	23.73	29.28
Narcissus mosaic virus (I)	6955	0.454	47.39	23.07	27.88	26.74	20.65	24.73
Clover yellow mosaic virus (I)	7015	0.440	49.42	22.80	31.82	30.04	19.39	18.76
Daphne virus-S (I)	8739	NC	45.10	NC	27.49	19.52	25.58	27.42
Papaya mosaic virus (I)	6656	0.748	47.93	23.91	30.18	25.20	22.73	21.89
Potato virus-M (I)	8635	0.655	48.54	25.45	26.46	20.11	28.44	24.98
Potato virus-X (I)	6435	0.685	46.79	22.48	30.66	23.84	22.95	22.55
White clover mosaic virus (I)	5845	0.730	44.09	22.53	30.27	27.72	16.37	25.65
Beet western yellows virus (I)	5646	NC	50.30	NC	27.50	25.80	24.50	22.20
Cymbidium mosaic virus (I)	6227	0.633	48.90	25.04	26.59	29.15	19.75	24.51
Luteoviridae								
Barley yellow dwarf virus-GAV (I)	5685	NC	48.04	NC	30.03	24.24	23.80	21.93
Potato leafroll virus (I)	5987	NC	49.46	NC	27.81	25.29	24.17	22.73
Sugarcane yellows leaf virus (I)	5899	NC	50.09	NC	26.72	26.06	24.04	23.02
Potyviridae								
Turnip mosaic virus (I)	9835	NC	45.65	21.96	31.88	21.28	24.37	22.47
Plum pox virus (I)	9741	0.802	43.41	21.30	31.34	20.45	22.96	25.24
Potato virus – Y (I)	9704	NC	42.15	NC	38.96	18.73	23.41	26.90
Tobacco etch virus (I)	9494	NC	43.19	NC	31.35	19.14	24.05	25.47
Benyvirus Group								
Beet necrotic yellows vein virus (+I)	6746	0.734	39.93	25.20	25.72	15.49	24.44	34.35

(*Contd.*)

Table 1. (*Contd.*)

Virus	Size (N)	CAI	GC%	GARP%	A%	C%	G%	T%
Furovirus Group								
Chinese wheat mosaic virus (II)	10716	0.597	43.76	20.65	28.00	17.63	26.13	28.25
Oat golden stripe virus (II)	10343	0.707	44.14	21.44	27.52	17.36	26.77	28.35
Soil borne cereal mosaic virus (II)	10708	0.744	43.52	20.64	28.05	17.02	26.50	28.43
Hordeivirus Group								
Barley stripe mosaic virus (III)	10221	0.622	42.58	22.44	28.77	19.20	23.38	28.65
Ideovirus Group								
Raspberry bushy dwarf virus (II)	7680	NC	42.94	NC	27.02	19.58	23.36	30.04
Tobravirus Group								
Pea early browning virus (II)	10447	NC	40.44	NC	29.90	15.99	24.46	29.65
Pepper ring spot virus (II)	8627	0.571	41.46	22.83	28.91	16.95	24.52	29.83
Tobacco rattle virus (II)	8805	NC	42.08	NC	28.94	16.60	25.47	28.98
Pecluvirus Group								
Indian peanut clump virus (II)	10338	0.720	43.53	22.07	26.66	18.40	25.13	29.81
Pomovirus Group								
Potato mop-top virus (III)	12141	0.714	42.83	22.50	28.61	16.95	25.88	28.56
Umbravirus Group								
Pea enation mosaic virus (I)	4223	0.576	55.91	32.53	22.41	27.38	28.36	21.68
Sobemovirus Group								
Southern bean mosaic virus (I)	4136	0.768	49.76	26.12	23.84	22.86	25.89	26.40
Tobamovirus Group								
Tobacco mosaic virus (I)	6384	0.719	41.71	19.96	29.86	28.56	23.15	28.43
Rhabdoviridae (-ssRNA viruses)								
Citurs psorosis virus (III)	11278	NC	34.64	NC	26.17	21.04	13.60	39.15
Lettuce necrotic yellows virus (II)	7868	NC	42.87	NC	31.24	18.86	24.01	25.89
Rice stripe virus (IV)	17145	0.360	38.79	17.09	32.73	18.75	20.03	28.48
Groundnut bud necrosis virus (II)	7858	0.431	34.88	16.25	32.57	17.80	17.08	32.55
Impatiens necrotic spot virus (I)	8776	0.342	32.82	14.19	29.69	18.55	14.27	37.49
Lettuce ring necrosis virus (IV)	12425	0.512	34.49	15.86	28.56	19.07	15.42	36.94
Mirafiore lettuce virus (IV)	12499	NC	34.51	16.47	28.39	19.27	15.24	37.02
Rice grassy stunt virus (VI)	25192	0.410	35.05	16.42	31.61	18.94	16.13	33.33
Watermelon silver mottle virus (III)	17381	0.381	34.05	14.34	34.72	16.27	17.78	31.23
Reoviridae (dsRNA viruses)								
Fiji disease virus (X)	29339	NC	31.91	NC	34.64	14.09	17.83	33.44
Rice ragged stunt virus (X)	26164	0.442	44.78	23.50	27.92	20.34	24.44	27.31
White clover cryptic virus (II)	3663	0.543	46.71	22.53	24.71	29.32	17.39	28.68
Lettuce big-vein virus (X)	6081	NC	45.19	NC	28.79	21.13	24.06	26.02

NC, means the CDS region of these genomes were not available.
Figures in parentheses indicate partite nature of virus genome, e.g. (I) means 'Monopartite'.

the influence of GC content on amino acid choice, we counted the number of amino acids Glycine, Alanine, Arginine and Proline (GARP), whose codons are GC-rich. The GARP contents in this set of viruses show a Pearson correlation coefficient of 0.959 ($P < 0.01$) with GC content. This indicates that amino acid content in the viral proteins is determined mainly by GC content of their respective genomes.

The CAI was designed for predicting the level of gene expression and assessing the adaptation of viral genes to their hosts. It is well known that highly expressed genes exhibit a strong bias for particular codons in many bacteria and small eukaryotes. One suggested explanation is the observation that there appears to be a relationship between tRNA abundance and codon bias[1].

Despite the importance of codon usage bias as an indicator of the forces shaping genome evolution, little is known about the extent and origin of this bias in RNA viruses. This is in contrast to organisms such as bacteria, yeast, *Drosophila* and mammals, where codon usage bias has been studied in much greater detail[9,10].

Codon usage bias may be the result of mutation pressure and/or natural selection for accurate and efficient translation. Mutation pressure has been shown to be the dominant factor shaping both codon usage bias and base composition in mammalian genomes[11,12] given that mutation rates in RNA viruses are much higher than those in life forms with DNA genomes[13]. Codon usage may also be influenced by an underlying bias in dinucleotide usage, for example, genes located in GC-rich regions of the chromosome preferentially utilize GC ending codons.

It is important for heterologous gene expression to encode proteins with sequences that yield optimal expression. A good thumb rule for finding such an optimal sequence is to choose codons that are most frequent in highly expressed genes. The CAI provides an explicit way of finding such codons; the most frequent codons simply have highest relative adaptiveness values, and sequences with higher CAIs are preferred over those with lower CAIs.

The study gives comprehensive information regarding the CAI and GC content of RNA genome plant viruses, and its influence on amino acid content.

1. Sharp, P. M., Stenico, M., Peden, J. F. and Lyod, A. T., *Biochem. Soc. Trans.*, 1993, **21**, 835–841.
2. Lobry, J. R. and Sueoka, N., *Genome Biol.*, 2002, **3**, 58.
3. Chen, S. L., Lee, W., Hottes, A. K., Shapiro, L. and McAdams, H. H., *Proc. Natl. Acad. Sci. USA*, 2004, **101**, 3480–3485.
4. Vartanian, J. P., Henry, M. and Wein-Hobson, S., *J. Gen. Virol.*, 2002, **83**, 801–805.
5. Aota, S. and Ikemura, T., *Nucleic Acids Res.*, 1986, **14**, 6345–6355.
6. Francino, M. P. and Ochman, H., *Nature*, 1999, **400**, 30–31.
7. Kanaya, S., Yamada, Y., Kinouchi, M., Kudo, Y. and Ikemura, T., *J. Mol. Evol.*, 2001, **53**, 290–298.
8. Haas, J., Park, E. C. and Seed, B., *Curr. Biol.*, 1996, **6**, 315–324.
9. Jansen, R., Bussemaker, H. J. and Gerstein, M., *Nucleic Acids Res.*, 2003, **31**, 2242–2251.
10. Mooers, A. O. and Holmes, E. C., *Trends Ecol. Evol.*, 2000, **15**, 365–369.
11. Wolfe, K., Sharp, P. M. and Li, W. H., *Nature*, 1989, **337**, 283–285.
12. Sharp, P. M. and Li, W. H., *Nucleic Acids Res.*, 1987, **15**, 1281–1295.
13. Drake, J. W. and Holland, J. J., *Proc. Natl. Acad. Sci. USA*, 1999, **96**, 13910–13913.

Received 20 March 2007; revised accepted 6 November 2007

U. S. KADAM[1,2,*]
S. B. GHOSH[1]

[1]Nuclear Agriculture and Biotechnology Division,
Bhabha Atomic Research Centre, Trombay,
Mumbai 400 085, India
[2]Present address: Department of Biotechnology,
National Research Centre for Grapes,
P.B. No. 03, Manjri Farm P.O.,
Solapur Road,
Pune 412 307, India
**For correspondence.*
e-mail: kadam_ulhas@yahoo.co.in

Antibacterial principles from the bark of *Terminalia arjuna*

The arjun tree *Terminalia arjuna* (Roxb.) is a well-known medicinal plant whose bark is extensively used in ayurvedic medicine, particularly as cardiac tonic. The bark is also prescribed in biliousness and sores and as an antidote to poison, and it is believed to have an ability to cure hepatic, congenital, venereal and viral diseases. A decoction of its bark with cane sugar and boiled cow's milk is highly recommended for endocarditis, pericarditis and angina[1].

Infectious endocarditis is an inflammatory disease of the endocardium, the internal lining of the human heart caused by bacteria such as staphylococci and gonococci. Among staphylococci, *Staphylococcus epidermidis* is one of the major etiological agents of this disease. The infections occur mainly in patients with prosthetic heart valves and during simple hospital procedures like catheterization, insertion of intra-uterine contraceptive devices, intravenous injections, etc.

In our screening programme aimed at detecting biomolecules from plant sources, which can specifically act against *S. epidermidis*, we found that the bark extracts of *T. arjuna* possessed antibacterial activity. Bioactivity-directed fractionation of the active extracts yielded three known oleane compounds: arjunic acid (**1**), arjungenin (**2**), and arjunetin (**3**), which were found to possess activity against *S. epidermidis*. The results presented here validate the traditional use of bark extracts of *T. arjuna* to cure endocarditis.

The bark of *T. arjuna* was collected from the CIMAP medicinal plants conservatory, during January 1999, identified in the Department of Botany and Pharmacognosy at CIMAP, where a voucher specimen (no. 5867) is maintained. The air-dried, powdered bark material was successively extracted with hexane and ethanol to yield hexane-soluble and alcohol-soluble fractions. The hexane and ethanol-insoluble plant material was extracted in water to get the water-soluble fraction. The alcohol-soluble extract was subsequently extracted with diethyl ether, ethyl acetate and methanol to yield the corresponding extracts.

For the isolation of pure molecules, *T. arjuna* (4.5 kg) was air-dried, crushed, powdered and extracted with hexane (3 × 5 l) at room temperature to remove fatty materials. The material was extracted with ethanol (3 × 5 l). The combined extract was concentrated under vacuum and further extracted using diethyl ether, which afforded 152 g of diethyl ether-soluble extract. The diethyl ether-soluble portion was column chromatographed over silica gel (60–120 mesh, 1200 g) using varying proportion of hexane : ethyl acetate (98 : 2, 95 : 5, 90 : 10, 85 : 15, 80 : 20, 75 : 25, 70 : 30, 60 : 40, 50 : 50, 40 : 60, 30 : 70, 20 : 80, 10 : 90, 100 : 0) as eluent. 100 ml of each fraction was collected and monitored by TLC.

Fraction nos 325–442 afforded compound **1**, identified as arjunic acid on the basis of spectral analysis[2–5], using hexane–ethyl acetate as eluent in the ratio (50 : 50 v : v) and crystallized using methanol.

Fraction nos 538–800 afforded compound **2**, identified as arjungenin by spectral analysis[3,6], using hexane–ethyl acetate as eluent in the ratio (50 : 50 v : v) and crystallized using methanol.

Fraction nos 949–1377 afforded compound **3**, identified as arjunetin by spectral analysis[3,7], using hexane–ethyl acetate as eluent in the ratio (20 : 80 v : v) and crystallized using methanol.