Horizontal gene transfer in herpesviruses identified by using support vector machine

M. FU[1], R. DENG[2], J. WANG[2], X. WANG[2*]

[1]Bioengineering and food science department, Guangdong University of Technology, Guangzhou, P.R. China; [2]State Key Laboratory of Biocontrol, School of Life Science, Sun Yat-Sen (Zhongshan) University, Guangzhou, P.R. China

Received December 21, 2010; accepted July 22, 2011

Summary. – Horizontal gene transfer (HGT) is the probable origin of new genes. Identification of HGT-introduced genes would be helpful to the understanding of the genome evolution and the function prediction of new genes. In this study, a method using support vector machine (SVM) was used to distinguish horizontally transferred genes and non-horizontally transferred genes of mammalian herpesviruses based on the atypical composition identification, with accuracy higher than 95% within a reasonable length of time by using just a common PC. This identified 302 putative horizontally transferred genes, 171 genes being identified for the first time. Although most putative transferred genes are of unknown function, many genes have been discovered or predicted to encode glycoproteins or membrane proteins.

Keywords: horizontal gene transfer; identification; herpesvirus

Introduction

HGT is one of the main mechanisms contributing to microbial genome evolution (Ochman *et al.*, 2000; Shackelton and Holmes, 2004), allowing rapid diversification and adaptation. It is also an important resource of new genes. Many "captured" new genes are biased to antibiotic resistance and pathogenicity-related function (Nakamura *et al.*, 2004; Willms *et al.*, 2006). Identification of genes introduced by HGT is important for the study of the genome evolution and the function prediction of new genes. In Lawrence and Ochman's studies (Lawrence and Ochman, 1997, 1998), three parameters, χ^2 of codon usage, the codon adaptation index CAI, and various indices of GC content, were used to evaluate HGT in *Escherichia coli*. In our previous paper, a novel method based on the frequencies of oligonucleotides was used to discover horizontally transferred genes in herpesvirus (Fu *et al.*, 2008). Here we present another novel method that exploits genomic composition to discover putative horizontally transferred genes in herpesviruses.

The herpesviruses are a group of large DNA viruses, which infect members of all groups of vertebrates, as well as some invertebrates. Herpesviruses have been typically classified into three subfamilies based upon biological and molecular characteristics. To date, eight discrete human herpesviruses have been described, each causing a characteristic disease. Herpesviruses have large genomes. One of them, cyprinid herpesvirus 3, has the largest genome, with approximately 300 kb of DNA encoding about one hundred and sixty genes. Among the proteins they encode, many have been distinguished to have essential viral functions, such as in genome replication and capsid assembly, or are being involved in direct interaction with the host, effecting immune evasion, cell proliferation, and apoptosis control. Many of these proteins are likely to have been acquired from the host to mimic or block normal cellular functions (Moore *et al.*, 1996; Alcami and Koszinowski, 2000; McFadden and Murphy, 2000). Identification and analysis of such "acquired" viral genes may lead to better understanding of the origin and evolution of these transferred genes and to

[*]Corresponding author. E-mail: wxz@mail.sysu.edu.cn; fax: +8620-84113964.

Abbreviations: COGs = cluster of orthologous groups; GCR = G protein-coupled receptor; HGT = horizontal gene transfer; SVM = support vector machine the development of therapeutic strategies to combat persistent viral infections.

Materials and Methods

General description. Each genome has a characteristic "signature", such as codon biases, short oligonucleotide composition and others, which is relatively constant throughout the genome (Mrazek and Karlin, 1999; Garcia-Vallvé *et al.*, 2000, 2003). Genes transferred from foreign organisms would retain the characteristic signature of its origin for a relatively long time (Lawrence and Ochman, 1997; Nakamura *et al.*, 2004). Hence, transferred genes can be detected on the basis of the signature difference between donor and receptor. This paper describes a distinct method that uses SVM to distinguish between transferred and non-transferred genes. SVM was trained using the signature of conserved mammalian herpesvirus genes as the features of the receptor, and that of conserved mammalian genes as the features of the donor. The method includes following steps: (1) collecting datasets, (2) generalizing the compositional features, and (3) training the SVM program and detecting the horizontally transferred genes.

Collection of datasets. Gene sequences of 33 mammalian herpesvirus genomes (Table 1) and the full-length gene sequences of four mammals (human, bovine, mouse and rat) were downloaded from Genbank. The conserved proteins were determined using the Tatusov method that was used to identify clusters of orthologous groups (COGs) in NCBI (Tatusov *et al.*, 1997, 2003). COGs of proteins were recognized by an all-against-all BLASP similarity search (Altschul *et al.*, 1997) among the 33 complete genomes. Herpesvirus genes were classified into COGs based on the protein sequence

Table 1. List of mammalian herpesvirus genomic sequences

Subfamily	Genus	Virus name	Acc. No.	ORFs	Abbre-viation	Length (kb)
Alpha-herpesvirinae	*Simplexvirus*	Cercopithecine herpesvirus 1	NC_004812	75	CeHV-1	157
		Cercopithecine herpesvirus 2	NC_006560	75	CeHV-2	151
		Human herpesvirus 1	NC_001806	77	HHV-1	152
		Human herpesvirus 2	NC_001798	77	HHV-2	155
	Varcellovirus	Bovine herpesvirus 1	NC_001847	73	BoHV-1	135
		Bovine herpesvirus 5	NC_005261	73	BoHV-5	138
		Cercopithecine herpesvirus 9	NC_002686	72	CeHV-9	124
		Equid herpesvirus 1	NC_001491	80	EHV-1	150
		Equid herpesvirus 4	NC_001844	79	EHV-4	146
		Human herpesvirus 3 (strain Dumas)	NC_001348	73	HHV-3	125
		Suid herpesvirus 1	NC_006151	77	SuHV-1	143
Beta-herpesvirinae	Cytomegalovirus	Cercopithecine herpesvirus 8	NC_006150	223	CeHV-8	221
		Human cytomegalovirus	NC_001347	151	HCMV	230
		Human herpesvirus 5 strain Merlin	NC_006273	165	HHV-5	236
		Pongine herpesvirus 4	NC_003521	165	PoHV-4	241
	Muromegalovirus	Murid herpesvirus 1	NC_004065	161	MuHV-1	230
		Murid herpesvirus 2	NC_002512	167	MuHV-2	230
	Roseolovirus	Human herpesvirus 6	NC_001664	123	HHV-6	159
		Human herpesvirus 6B	NC_000898	104	HHV-6B	162
		Human herpesvirus 7	NC_001716	86	HHV-7	153
	Unclassified	Tupaiid herpesvirus 1	NC_002794	158	TuHV-1	196
Gamma-herpesvirinae	Lymphocryptovirus	Callitrichine herpesvirus 3	NC_004367	72	CalHV-3	150
		Cercopithecine herpesvirus 15	NC_006146	80	CeHV-15	171
		Human herpesvirus 4	NC_001345	94	HHV-4	172
	Rhadinovirus	Alcelaphine herpesvirus 1	NC_002531	71	AlHV-1	131
		Bovine herpesvirus 4	NC_002665	79	BoHV-4	109
		Cercopithecine herpesvirus 17	NC_003401	89	CeHV-17	134
		Equid herpesvirus 2	NC_001650	79	EHV-2	184
		Human herpesvirus 8	NC_003409	82	HHV-8	138
		Murid herpesvirus 4	NC_001826	81	MuHV-4	119
		Saimiriine herpesvirus 2	NC_001350	76	SaHV-2	113
Unclassified		Ateline herpesvirus 3	NC_001987	73	AtHV-3	108
		Macaca fuscata rhadinovirus	NC_007016	171	MFRV	131

similarity. Homology was determined by the significance of the BLAST hit and by the length of the maximal scoring pair alignment in the BLASTP search. Two proteins were deemed homologous if bidirectional BLASTP hits (score ≥50) produced alignments that covered at least 50% of the query sequence with e-value $\leq 10^{-4}$. Given pairwise hits among herpesvirus proteins, COGs were defined by single-linkage clustering. 20 groups of proteins (660 genes, listed in Table 2) that were conserved in 33 herpesviruses were used as the dataset of non-transferred genes. The same number of mammalian conserved genes identified with the same method was used as the dataset of transferred genes for module training. All genes, excluding the 660 conserved genes in herpesviruses, were used as the test dataset, and the same number of genes from the four mammals (other than the 660 conserved genes) were selected and used as the negative control dataset.

Generalization of compositional features. The compositional feature vector for any given DNA sequence over a set of templates $\pi=\{\pi_1, \pi_2,\ldots, \pi_q\}$ is denoted as $\Phi(s)=(p_1, p_2,\ldots, p_q)$; here π_i is k-mer oligonucleotide template $\alpha_1\alpha_2\ldots\alpha_k$, p_i is the frequency of template π_i in sequence. Instead of using the absolute template frequencies, we normalize these frequencies over the expected template frequencies, which can be derived from the single nucleotide composition:

$$P = \frac{p(\alpha_1\alpha_2\cdots\alpha_k)}{\prod_{j=1}^{k} p(\alpha_j)}$$

where $p(\alpha_1\alpha_2\ldots\alpha_k)$ is the frequency of template $\alpha_1\alpha_2\ldots\alpha_k$, $p(\alpha_j)$ is the frequency of the jth nucleotide of the template $\alpha_1\alpha_2\ldots\alpha_k$. So every gene is depicted by a 4^k dimension vector.

Training of the svm_learn.exe in the SVMlight and discrimination of the transferred genes by svm_classify.exe in SVMlight. SVM*light* (Joachims, 1998), including svm_learn.exe and svm_classify.exe, is an implementation of SVMs in C. SVM is a new pattern-recognition method based on recent advances in statistical learning theory. Given two sets of training data points in a high-dimensional input space, the objective of the SVM method is to learn a function that will take the value of +1 in the region, where the positive data points are concentrated, and the value of −1, where the negative points are concentrated. The function to be learned is modeled as a hyperplane in a transformed space (=feature space), and hyperplane parameters are estimated so that its margin with respect to the training data is maximized.

The compositional features calculated from the training datasets were used to train the learning module svm_learn.exe with the linear kernel function and cost factor 1. The ideal output of positive data and negative data were set to +1 and −1, respectively. The training result was used as a classifier input file for the classify module svm_classify.exe to discriminate between transferred and non-transferred genes in the test dataset. The negative control dataset was used to test the accuracy of discrimination.

Table 2. List of herpesvirus conserved gene families

Acc. No. (HHV-1)	Gene name (HHV-1)	Function[a]	Functional class[b]
GI:9629382	UL2	uracil-DNA glycosylase	Nuc
GI:9629385	UL5	component of DNA helicase-primase complex	Rep
GI:9629386	UL6	minor capsid protein	Str
GI:9629387	UL7	unknown	Unk
GI:9629390	UL10	virion glycoprotein M	Gly
GI:9629392	UL12	deoxyribonuclease	Nuc
GI:9629393	UL13	protein kinase	Oth
GI:9629398	UL18	capsid protein	Str
GI:9629402	UL22	virion glycoprotein H	Gly
GI:9629404	UL24	unknown	Unk
GI:9629405	UL25	capsid associated tegument protein	Str
GI:9629406	UL26	protease	Str
GI:9629408	UL27	virion glycoprotein B	Gly
GI:9629409	UL28	DNA packaging	Str
GI:9629412	UL31	unknown	Unk
GI:9629413	UL32	virion protein	Str
GI:9629420	UL39	ribonucleotide reductase large subunit	Nuc
GI:9629432	UL50	deoxyuridine triphosphatase	Rep
GI:9629434	UL52	component of DNA helicase-primase complex	Rep
GI:9629436	UL54	immediate early protein	Trf

[a] Function as derived from GenBank annotations. [b] Functional classes: Rep (replication), Nuc (nucleotide metabolism and DNA repair), Str (structural), Trf (transcription), Gly (glycoprotein), Oth (other), Unk (unknown).

Table 3. The accuracy (%) of discrimination based on SVM

	k=2		k=3		k=4	
	T[a]	F[b]	T[a]	F[b]	T[a]	F[b]
human	96.7	3.3	94.4	5.6	94.0	6.0
mouse	95.8	4.2	94.6	5.4	95.2	4.8
rat	96.1	3.9	95.8	4.2	95.1	4.9
bovine	94.6	5.4	94.8	5.2	94.1	5.9
average	95.8		94.9		94.6	

[a]T: correct discrimination; [b]F: incorrect discrimination.

Results

Datasets and their composition features

The k-mer oligonucleotide frequencies of the 660 conserved genes in herpesviruses and the 660 conserved genes in the four mammals were used as the positive dataset and the negative dataset (non-transferred genes and transferred genes), respectively. The frequencies of 2721 genes other than the conserved genes in herpesviruses were taken as the test dataset, and those of 1143 genes other than the conserved genes for every mammal were taken as the negative control dataset (the maximum number of the bovine full length genes other than the conserved genes in the database was 1143, so the same number of other mammalian genes was used). The number of data for every gene was dependent on the number of the nucleotides in the template, which was 4^k in k-mer oligonucleotide frequency.

Training of the learning module and discrimination for the transferred genes

The output of each training procedure was a classifier file, which was then used for the classify module to discriminate the test data. Using this new method, altogether 302 genes in herpesviruses (length ≥500bp) were detected as putative transferred genes (Table 4), among which 266 were from Gammaherpesviruses, 32 from Betaherpesviruses, and 4 from Alphaherpesviruses, implying that gene acquisition in Gammaherpesvirus was more active than in the other two groups, which agreed with Holzerlandt's results (Holzerlandt *et al.*, 2002) and our previous conclusion (Fu *et al.*, 2008). Although most putative transferred genes are of unknown function, many genes have been discovered or predicted as encoding glycoproteins or membrane proteins.

Discrimination accuracy

The result of transferred gene discrimination (shown in Table 3) indicated that discrimination accuracy reached more than 90% when normalized frequencies of k-mer oligonucleotide (k = 2, 3, or 4) were taken as the compositional feature vectors to quantify the genes. Comparative evaluation of different methods for quantifying genes demonstrated that using dinucleotides as a template already has yielded the best discrimination result, which only required to process 4^2 data for every gene, computer resource affordable using even a general private computer.

Table 4. Horizontally transferred genes predicted by the method based on SVM

Abbreviation	Gene name	Description[a]	Acc. No. (Gi)	Length (bp)
CeHV-8	rh31 rh31*	similar to human cytomegalovirus UL13	GI:51556491	1307
CeHV-8	rh167*		GI:51556622	503
CeHV-8	rh224		GI:51556677	617
HCMV	UL122	IE2; immediate-early transcriptional regulator;	GI:28373220	3401
HCMV	UL123	IE1; immediate-early transcriptional regulator;	GI:9625811	1759
HHV-5	UL122*	IE2; immediate-early transcriptional regulator;	GI:52139286	3401
HHV-5	UL123*	IE1; immediate-early transcriptional regulator;	GI:52139287	1761
PoHV-4	tegument protein UL71	similar to HSV-1 UL51	GI:20026662	1121
PoHV-4	immediate-early transcriptional regulator UL122	IE2	GI:20026703	3521
PoHV-4	immediate-early transcriptional regulator UL123	IE1	GI:20026704	1806
PoHV-4	glycoprotein US6*	inhibits TAP-mediated peptide translocation; US6 family	GI:20026737	578
PoHV-4	US19	contains multiple hydrophobic regions; US12 family	GI:20026750	818
PoHV-4	US34*		GI:20026763	509
CalHV-3	ORF6*	similar to EBV BILF1; glycoprotein gp64; GCR (Paulsen *et al.*, 2005)	GI:24943096	917
CalHV-3	ORF19	similar to EBV BDLF2	GI:24943110	1175

Table 4 continued

CalHV-3	ORF21	similar to EBV BDLF4	GI:24943112	635
CalHV-3	ORF32*	similar to EBV BBLF3; helicase-primase complex	GI:24943123	602
CalHV-3	ORF41	similar to EBV BRRF1	GI:24943132	911
CalHV-3	C3		GI:24943136	2569
CalHV-3	C4*		GI:24943157	764
CeHV-15	BCRF1*	similar to Epstein-Barr virus BCRF1; viral interleukin-10 (Rivailler *et al.*, 2002)	GI:51518017	533
CeHV-15	EBNA-LP	similar to Epstein-Barr virus EBNA-LP	GI:51518018	18102
CeHV-15	BHRF1*	similar to Epstein-Barr virus BHRF1; bcl-2 homologue (Rivailler *et al.*, 2002)	GI:51518020	575
CeHV-15	BFRF1	similar to Epstein-Barr virus BFRF1; tegument (Rivailler *et al.*, 2002)	GI:51518023	989
CeHV-15	BFRF2	similar to Epstein-Barr virus BFRF2	GI:51518024	1811
CeHV-15	BFRF3	similar to Epstein-Barr virus BFRF3; capsid protein (Rivailler *et al.*, 2002)	GI:51518094	512
CeHV-15	BaRF1*	similar to Epstein-Barr virus BaRF1; Ribonucleotide reductase (Rivailler et al, 2002)	GI:51518029	908
CeHV-15	BMRF1*	similar to Epstein-Barr virus BMRF1; dsDNA binding protein	GI:51518030	1214
CeHV-15	BMRF2	similar to Epstein-Barr virus BMRF2; Membrane protein (Rivailler *et al.*, 2002)	GI:51518031	1073
CeHV-15	BSRF1*	similar to Epstein-Barr virus BSRF1; tegument	GI:51518034	665
CeHV-15	EBNA-3A*	similar to Epstein-Barr virus EBNA-3A; latent infection nuclear proteins important for Epstein-Barr virus (EBV)-induced B-cell immortalization and the immune response to EBV infection. (Jiang *et al.*, 2000)	GI:51518092	2912
CeHV-15	EBNA-3B*	similar to Epstein-Barr virus EBNA-3B; latent infection nuclear proteins important for *Epstein-Barr virus* (EBV)-induced B-cell immortalization and the immune response to EBV infection. (Jiang *et al.*, 2000)	GI:51518091	2867
CeHV-15	BZLF2	similar to Epstein-Barr virus BZLF2; Glycoprotein, gp42	GI:51518040	665
CeHV-15	BZLF1*	similar to Epstein-Barr virus BZLF1; Transactivator (Rivailler *et al.*, 2002)	GI:51518041	1096
CeHV-15	BRLF1*	similar to Epstein-Barr virus BRLF1; Transactivator (Rivailler *et al.*, 2002)	GI:51518042	1808
CeHV-15	BRRF1*	similar to Epstein-Barr virus BRRF1	GI:51518043	929
CeHV-15	BRRF2	similar to Epstein-Barr virus BRRF2	GI:51518044	1505
CeHV-15	EBNA-1*	similar to Epstein-Barr virus EBNA-1; sequence-specific DNA-binding proteins (Johannsen *et al.*, 2004)	GI:51518045	1535
CeHV-15	BKRF4*	similar to Epstein-Barr virus BKRF4; tegument protein (Johannsen *et al.* 2004)	GI:51518048	719
CeHV-15	BDRF1	similar to Epstein-Barr virus BDRF1; Packaging protein (Rivailler *et al.*, 2002)	GI:51518058	5824
CeHV-15	BDLF4	similar to Epstein-Barr virus BDLF4	GI:51518062	716
CeHV-15	BDLF3*	similar to Epstein-Barr virus BDLF3; envelope glycoprotein (Johannsen *et al.*, 2004)	GI:51518063	779
CeHV-15	BcLF1*	similar to Epstein-Barr virus BcLF1; capsid protein (Johannsen *et al.*, 2004)	GI:51518066	4142
CeHV-15	BcRF1	similar to Epstein-Barr virus BcRF1	GI:51518067	1733
CeHV-15	BTRF1*	similar to Epstein-Barr virus BTRF1	GI:51518068	1211
CeHV-15	BXLF1*	similar to Epstein-Barr virus BXLF1; thymidine kinase (Johannsen *et al.*, 2004)	GI:51518070	1823
CeHV-15	LF3	similar to Epstein-Barr virus LF3	GI:51518075	2672
CeHV-15	BILF1*	similar to Epstein-Barr virus BILF1; GCR (Paulsen *et al.*, 2005)	GI:51518078	938
CeHV-15	BALF5*	similar to Epstein-Barr virus BALF5; DNA polymerase (Rivailler *et al.*, 2002)	GI:51518080	3047
CeHV-15	ECRF4	similar to Epstein-Barr virus ECRF4	GI:51518079	1136
CeHV-15	BARF1*	similar to Epstein-Barr virus BARF1; CSF-1R (Rivailler *et al.*, 2002)	GI:51518086	662
CeHV-15	LMP1*	similar to Epstein-Barr virus LMP1	GI:51518089	1939
HHV-4	unnamed protein product*	BNRF1 reading frame; major tegument protein; vFGAM	GI: 9625579	3956

Table 4 continued

HHV-4	unnamed protein product*	BCRF1 reading frame; vIL-10	GI: 9625580	512
HHV-4	unnamed protein product	BCRF2 reading frame 1	GI: 9625581	1151
HHV-4	unnamed protein product	BWRF1 reading frame 2	GI:9625582	1151
HHV-4	unnamed protein product	BWRF1 reading frame 3	GI:9625583	1151
HHV-4	unnamed protein product	BWRF1 reading frame 4	GI:9625584	1151
HHV-4	unnamed protein product	BWRF1 reading frame 5	GI:9625585	1151
HHV-4	unnamed protein product	BWRF1 reading frame 6	GI:9625586	1151
HHV-4	unnamed protein product	BWRF1 reading frame 7	GI:9625587	1151
HHV-4	unnamed protein product	BWRF1 reading frame 8	GI:9625588	1151
HHV-4	unnamed protein product	BWRF1 reading frame 9	GI:9625589	1151
HHV-4	unnamed protein product	BWRF1 reading frame 10	GI:9625590	1151
HHV-4	unnamed protein product	BWRF1 reading frame 11	GI:9625591	1151
HHV-4	unnamed protein product	BWRF1 reading frame 12	GI:9625592	1151
HHV-4	unnamed protein product	BFRF2 early reading frame, homologous to HFLF5 in CMV	GI:9625597	1775
HHV-4	unnamed protein*	BPLF1 reading frame; Tegument protein	GI: 9625599	9449
HHV-4	unnamed protein product*	BaRF1 early reading frame, Ribonucleotide reductase, small subunit	GI: 9625603	908
HHV-4	unnamed protein product*	BMRF1 early reading frame. Early antigen protein recognised by R3 monoclonal	GI: 9625604	1214
HHV-4	unnamed protein product	BMRF2 early reading frame. Membrane protein (Rivailler *et al.*, 2002)	GI:9625605	1073
HHV-4	unnamed protein product	BSRF1 reading frame	GI:9625609	656
HHV-4	unnamed protein product*	EBNA3B (EBNA4A); latent infection nuclear proteins important for *Epstein-Barr virus* (EBV)-induced B-cell immortalization and the immune response to EBV infection. (Jiang *et al.*, 2000)	GI: 9625617	2894
HHV-4	unnamed protein product	EBNA3C (EBNA 4B) latent protein (Jiang *et al.*, 2000)	GI:9625618	3052
HHV-4	unnamed protein product	BZLF1 reading frame; Transactivator	GI:9625620	945
HHV-4	unnamed protein product*	BRLF1 reading frame, (immediate?) early gene, acts as transcription activator	GI: 9625622	1817
HHV-4	unnamed protein product	BRRF1 early reading frame	GI:9625621	932
HHV-4	unnamed protein product	BRRF2 reading frame	GI:9625623	1613
HHV-4	unnamed protein product*	BKRF1 encodes EBNA-1 protein, latent cycle gene	GI: 9625624	1925
HHV-4	unnamed protein product*	BKRF4 reading frame, contains complex repetitive sequence	GI: 9625627	653
HHV-4	unnamed protein product	BBLF3 early reading frame, spliced to BBLF2. BBLF3 contains a consensus nucleotide binding site; Helicase-primase complex (Rivailler *et al,* 2002)	GI:9625631	602
HHV-4	unnamed protein product	BGLF3 reading frame	GI:9625638	998

Table 4 continued

HHV-4	unnamed protein product	probable DNA packaging protein; BDRF1 reading frame	GI:9625642	5413
HHV-4	unnamed protein product	BGRF1 reading frame, Packaging protein (Rivailler et al, 2002)	GI:9625637	977
HHV-4	unnamed protein product	BGLF1 late reading frame	GI:9625640	1523
HHV-4	unnamed protein product	BDLF4 early reading frame	GI:9625641	677
HHV-4	unnamed protein product*	BDLF2 late reading frame; tegument	GI: 9625644	1262
HHV-4	unnamed protein product	BcRF1 reading frame		1727
HHV-4	unnamed protein product	BTRF1 reading frame. Northern blots detect 0.95 late and 3.8kb early RNA		1274
HHV-4	unnamed protein product*	BXLF1 early reading frame, thymidine kinase.	GI: 9625651	1823
HHV-4	unnamed protein product*	BILF1 reading frame, membrane protein, 3xNXS /T; GCR (Paulsen *et al.*, 2005)	GI: 9625656	938
HHV-4	unnamed protein product*	BALF5 DNA polymerase (early), homologous to many DNA polymerases, CMV HFLF2 and RF 28 VZV. 4.5kb early RNA apparently encodes BALF5, RNA ends unknown	GI:9625657	3047
HHV-4	unnamed protein product*	BARF1 reading frame a secretory protein with transforming and mitogenic activities (Wang *et al.*, 2006); CSF-1R (Rivailler *et al.*, 2002)	GI: 9625661	665
MHV-2	pR122-EX5		GI:9845417	1517
MHV-2	pR123-EX3		GI:9845419	2611
MHV-2	pR123-EX4		GI:9845418	1295
MHV-2	pr128	US22 family homolog;	GI:9845425	1226
AlHV-1	A3	semaphorin homolog; AHV-sema, similar to Vaccinia A39	GI:10140929	1961
AlHV-1	ORF03*	tegument protein; similar to H. saimiri and EHV2 ORF3, similar to ORF75; Virion protein, FGARAT (Ensser *et al.*, 1997)	GI:10140931	4109
AlHV-1	Putative BALF1 homolog*	Putative antagonist of herpesvirus BCL-2;	GI:19343407	695
AlHV-1	ORF06	major ss DNA binding protein	GI:10140932	3383
AlHV-1	ORF09*	DNA Polymerase; similar to EBV BALF5, CMV UL54, HSV UL30	GI:10140935	3080
AlHV-1	A5*	similar to EBV BILF1; possible seven transmembrane protein with similarity to G-protein coupled receptors	GI:10140936	908
AlHV-1	ORF10*	similar to H. saimiri, EHV2, KSHV ORF10, EBV Raji LF1	GI:10140937	1214
AlHV-1	ORF18	similar to CMV UL 79	GI:10140940	827
AlHV-1	ORF 23	similar to EBV BTRF1	GI:10140945	1205
AlHV-1	ORF25	major capsid protein; ORF25; similar to EBV BCLF1, CMV UL75, HSV UL22	GI:10140947	4112
AlHV-1	ORF33	similar to EBV BGLF2, CMV UL94, HSV UL16	GI:10140954	1007
AlHV-1	ORF34	similar to EBV BGLF3, CMV UL95, HSV UL14	GI:10140955	1031
AlHV-1	ORF45	similar to EBV BKRF4	GI:10140966	707
AlHV-1	ORF47	similar to CMV UL115 gL, HSV UL1 and EBV BKRF2; weak positional homologue	GI:10140968	506
AlHV-1	ORF48	similar to EBV BRRF2	GI:10140969	1259
AlHV-1	ORF50*	R-transactivator; similar to EBV BRLF1; splicing predicted by splice site analysis	GI:10140970	2078
AlHV-1	A6*	position similar to EBV BZLF1; Transactivator (Rivailler *et al.*, 2002)	GI:10140971	632
AlHV-1	ORF55	similar to EBV BSRF1, CMV UL71, HSV UL51	GI:10140977	662
AlHV-1	ORF58	similar to EBV BMRF2 and HSV UL43	GI:10140980	1055
AlHV-1	ORF59*	processivity factor; DNA replication; subunit of DNA-polymerase; similar to EBV BMRF1 (EA-D), CMV UL44, HSV UL42	GI:10140981	1235
AlHV-1	ORf60*	ribonucleotide-reductase, small subunit; RRsmall; similar to EBV BARF1, HSV UL40	GI:10140982	917
AlHV-1	ORF63	tegument protein; similar to EBV BOLF1, CMV UL47, HSV UL37	GI:10140985	2858
AlHV-1	ORF64	large tegument protein; similar to EBV BPLF1, CMV UL48, HSV UL36	GI:10140986	7820

Table 4 continued

AlHV-1	ORF65	capsid protein; positional similar to EBV BFRF3 and HSV UL35	GI:10140987	758
AlHV-1	ORF66	similar to EBV BFRF2, CMV UL39	GI:10140988	1313
AlHV-1	ORF67	tegument protein; virion tegument protein; similar to EBV BFRF1, CMV UL50, HSV UL34	GI:10140989	791
AlHV-1	ORF73*	putative immediate early protein; similar to H. saimiri and KSHV ORF73	GI:10140993	3902
AlHV-1	ORF75*	similar to ORF3 of EHV2 and AHV-1, and ORF75 of all rhadinoviruses, and EBV BNRF1; also similar to formylglycineamide-synthase	GI:10140994	3947
AlHV-1	A9*	similar to Bcl-family of proteins; contains only conserved BH1 domain; functional similarity may exist to ORF16 of H. saimiri, KSHV, BHV4 and EBV BHRF1	GI:10140995	506
AlHV-1	A10	putative glycoprotein	GI:10140996	1418
BoHV-4	ORF3 BORFA1*	v-FGAM-synthase; tegument protein;	GI:13095580	3866
BoHV-4	ORF 6*	single-stranded DNA-binding protein MDBP	GI:13095583	3404
BoHV-4	ORF 9*	DNA polymerase;	GI:13095586	3017
BoHV-4	ORF 10	BORFB1; conserved in other gamma-herpesviruses	GI:13095587	1280
BoHV-4	pBo5	hypothetical protein; long ORF of immediate early transcript 1 RNA, exons I-IV	GI:13095589	1139
BoHV-4	ORF 16*	BORFB2; v-Bcl-2 protein	GI:13095593	680
BoHV-4	ORF 21*	thymidine kinase	GI:13095598	1337
BoHV-4	ORF 23	conserved in other gamma-herpesviruses	GI:13095600	1202
BoHV-4	ORF 24	conserved in other herpesviruses	GI:13095601	2258
BoHV-4	ORF 25	major capsid protein	GI:13095602	4121
BoHV-4	ORF 27	conserved in other gamma-herpesviruses	GI:13095604	638
BoHV-4	ORF 29	cleavage/packaging protein; exons I and II	GI:13095610	5174
BoHV-4	ORF 31	conserved in other gamma-herpesviruses	GI:13095607	641
BoHV-4	ORF 32	viral DNA cleavage/packaging protein	GI:13095608	1370
BoHV-4	ORF 33	conserved in other herpesviruses	GI:13095609	998
BoHV-4	ORF 34	conserved in other herpesviruses	GI:13095611	986
BoHV-4	ORF 40	helicase-primase complex component	GI:13095617	1373
BoHV-4	ORF 41	helicase-primase complex component	GI:13095618	521
BoHV-4	ORF 45*	unknown	GI:13095622	725
BoHV-4	ORF 48	conserved in other gamma-herpesviruses	GI:13095625	1544
BoHV-4	ORF 50	R transactivator protein; exons I and II; encoded by immediate early transcript 2 RNA	GI:13095626	2593
BoHV-4	ORF 49*	unknown	GI:13095627	899
BoHV-4	ORF 55*	unknown	GI:13095632	602
BoHV-4	ORF 58	conserved in other gamma-herpesviruses	GI:13095635	1052
BoHV-4	ORF 59	DNA replication protein	GI:13095636	1175
BoHV-4	ORF 60*	ribonucleotide reductase small subunit	GI:13095637	917
BoHV-4	ORF 62	assembly/DNA maturation protein	GI:13095639	1019
BoHV-4	ORF 63	tegument protein	GI:13095640	2819
BoHV-4	ORF 64	tegument protein	GI:13095641	7709
BoHV-4	ORF 66	conserved in other herpesviruses	GI:13095643	1274
BoHV-4	ORF 67	tegument protein	GI:13095644	770
BoHV-4	ORF 71	v-FLIP; BORFE2	GI:13095651	548
BoHV-4	ORF 75*	tegument protein/v-FGAM-synthetase	GI:13095653	3917
BoHV-4	ORF Bo14*	BORFF1; hypothetical protein pBo14; proline rich (Zimmermann *et al.*, 2001)	GI:13095654	512
BoHV-4	ORF Bo17*	BORFF3-4; v-beta-1,6GnT	GI:13095657	1322
EHV-2	ORF E4*		GI:9628007	551
EHV-2	ORF E6*	putative GCR	GI:9628013	977
EHV-2	ORF 17.5	capsid scaffold protein	GI:9628018	1010
EHV-2	ORF 21*	thymidine kinase	GI:9628023	1841
EHV-2	ORF 33		GI:9628035	1025
EHV-2	ORF 45		GI:9628048	965
EHV-2	ORF 48	glycoprotein L	GI:9628051	1832

Table 4 continued

EHV-2	ORF 50	transcriptional control	GI:9628053	1892
EHV-2	ORF 55*		GI:9628058	674
EHV-2	ORF 59*	DNA polymerase processivity subunit	GI:9628062	1244
EHV-2	ORF 62	capsid protein; intercapsomeric triplex	GI:9628065	1016
EHV-2	ORF 63	tegument protein	GI:9628066	2897
EHV-2	ORF 64	tegument protein	GI:9628067	10310
EHV-2	ORF 65	capsid protein; external to capsomers	GI:9628068	539
EHV-2	ORF 66		GI:9628069	1370
EHV-2	ORF 67	tegument protein	GI:9628070	863
EHV-2	ORF E7*	interleukin 10-like protein, similar to protein encoded by GenBank Accession Number S59624	GI:9628072	539
EHV-2	ORF 70*	thymidylate synthase	GI:9628075	869
EHV-2	ORF 74*	GCR	GI:9628076	992
EHV-2	ORF E8*		GI:9628077	515
EHV-2	ORF 75	tegument protein	GI:9628078	4037
EHV-2	ORF E10		GI:9628080	632
MFRV	JM145		GI:66476694	869
MHV-4	M1	serpin	GI:9629554	1262
MHV-4	M2*		GI:9629599	599
MHV-4	M3*		GI:9629600	1220
MHV-4	M4*	GCR homologue	GI:9629555	1379
MHV-4	ORF4	complement regulatory protein	GI:9629556	1166
MHV-4	ORF6*	ssDNA binding protein	GI:9629557	3311
MHV-4	ORF9*	DNA polymerase	GI:9629560	3083
MHV-4	ORF10		GI:9629561	1256
MHV-4	ORF11*		GI:9629562	1166
MHV-4	K3	BHV4-IE1 homolog	GI:9629601	605
MHV-4	M6		GI:9629564	1757
MHV-4	ORF18b		GI:9629565	854
MHV-4	ORF21*	thymidine kinase	GI:9629566	1934
MHV-4	ORF23*		GI:9629605	1145
MHV-4	ORF24		GI:9629606	2153
MHV-4	ORF25	major capsid protein	GI:9629568	4121
MHV-4	ORF27*		GI:9629570	764
MHV-4	ORF29b	packaging protein	GI:9629607	1046
MHV-4	ORF31		GI:9629572	602
MHV-4	ORF32*		GI:9629573	1334
MHV-4	ORF33*		GI:9629574	983
MHV-4	ORF29a	packaging protein	GI:9629608	920
MHV-4	ORF34		GI:9629575	998
MHV-4	ORF40*	helicase-primase	GI:9629580	1832
MHV-4	ORF45*		GI:9629612	620
MHV-4	ORF47*	glycoprotein L	GI:9629614	521
MHV-4	ORF48*		GI:9629615	1001
MHV-4	ORF49		GI:9629616	905
MHV-4	ORF50	transcriptional activator	GI:9629582	1469
MHV-4	M7*	glycoprotein 150	GI:9629583	1451
MHV-4	ORF55*		GI:9629619	572
MHV-4	ORF58		GI:9629620	1043
MHV-4	ORF59*	DNA replication protein	GI:9629621	1184
MHV-4	ORF60*	ribonucleotide reductase small subunit	GI:9629622	917
MHV-4	ORF62	assembly/DNA maturation	GI:9629624	1142
MHV-4	ORF63	tegument protein	GI:9629588	2816
MHV-4	ORF64	tegument protein	GI:9629589	7373
MHV-4	M9*		GI:9629625	560
MHV-4	ORF66		GI:9629626	1229
MHV-4	ORF67	tegument protein	GI:9629627	680

Table 4 continued

MHV-4	M10a		GI:9629592	2324
MHV-4	M10b		GI:9629593	2120
MHV-4	ORF72*	cyclin D homolog	GI:9629628	758
MHV-4	M11*	bcl-2 homolog	GI:9629595	515
MHV-4	ORF73	immediate-early protein	GI:9629629	944
MHV-4	ORF74*	GCR (IL8 receptor homolog?)	GI:9629596	1013
MHV-4	ORF75C*	tegument protein G75C	GI:9629630	3932
MHV-4	ORF75B*	tegument protein G75B	GI:9629631	3827
MHV-4	ORF75A*	tegument protein G75A	GI:9629632	3875
MHV-4	M12		GI:9629597	692
MHV-4	M13		GI:9629598	638
SaHV-2	ORF 02*	dihydrofolate reductase	GI:9625957	563
SaHV-2	ORF 03*	similarity to ORF 75 and EBV BNRF1	GI:9625958	3740
SaHV-2	Orf 09 KCRF2*	DNA polymerase	GI:9625965	3029
SaHV-2	ORF 10 KCRF3		GI:9625966	1223
SaHV-2	ORF 12 KCLF1		GI:9625968	509
SaHV-2	ORF 23	similar to EBV BTRF1	GI:9625979	761
SaHV-2	ORF 25	major capsid protein	GI:9625981	4115
SaHV-2	ORF 33	similar to other herpesviruses	GI:9625988	992
SaHV-2	ORF 34	similar to other herpesviruses	GI:9625989	950
SaHV-2	ORF 40	similar to EBV BBLF2	GI:9625996	1352
SaHV-2	ORF 45*	similar to EBV BKRF4	GI:9626001	773
SaHV-2	ORF 48	EDLF5 similar to EBV BRRF2	GI:9626004	2393
SaHV-2	ORF 49 EDLF4*	similar to EBV BRRF1	GI:9626006	911
SaHV-2	ORF 50 EDRF1	Herpesvirus S.R transactivator; sim. to EBV BRLF1, putative 3’-ORF, 5’-exon unknown”	GI:9626005	1607
SaHV-2	ORF 55 EDLF1*	similar to EBV BSRF1	GI:9626012	602
SaHV-2	ORF 59 EELF4	similar to EBV BMRF1	GI:9626015	1106
SaHV-2	ORF 60 EELF3*	ribonucleotide reductase, small subunit	GI:9626016	917
SaHV-2	ORF 62	EELF1 similar to other herpesviruses	GI:9626018	992
SaHV-2	ORF 64	EERF2 similar to other herpesviruses	GI:9626020	7409
SaHV-2	ORF 70 ECLF4*	thymidylate synthase	GI:9626026	884
SaHV-2	ORF 71 ECLF3		GI:9626027	503
SaHV-2	ORF 72 ECLF2*	cyclin homologue	GI:9626028	764
SaHV-2	ORF 73 ECLF1*		GI:9626029	1223
SaHV-2	ORF 75*	EILF1 similar to ORF 03 and EBV BNRF1	GI:9626031	3899
HHV-6	U86, IE2	Transactivation; old name BCLF1; homologue HCMV UL122, IE2; region IE-A, immediate early gene	GI:9628388	2147
HHV-6	U87	possible glycoprotein; region IE-B, highly charged, pro repeats; presenting U86 /U87 as one ORF, BCLF0	GI:9628389	2492
HHV-6	U89*	Transactiviation, IE1	GI:9628391	2519
HHV-6B	U44	major immediate-early protein; IE-A	GI:9633155	4562
HHV-6B	U45*	IE-A transactivator	GI:9633156	3433
HHV-7	U7*	betaherpesvirus US22 gene family; exons 1 and 2 are similar to HHV-5 UL28/UL29; exon 3 related to HHV-7 U4 and similar to HHV-5 UL27	GI:89112557	3855
HHV-7	U25	betaherpesvirus US22 gene family; similar to HHV-5 UL43	GI:51874247	962
HHV-7	U30	tegument protein; herpesvirus core gene UL37 family; similar to HHV-5 UL47	GI:51874252	2816
HHV-7	U45	herpesvirus core gene UL50 family; herpesvirus DURP gene family; similar to HHV-5 UL72; related to dUTPase but probably not enzymatically active	GI:51874267	1139
HHV-7	U52*	similar to HHV-5 UL79	GI:51874274	764
HHV-7	U53.5*	major capsid scaffold protein; herpesvirus core gene UL26.5 family; similar to HHV-5 UL80.5	GI:51874276	692
HHV-7	U86	IE-A protein; similar to HHV-5 UL122	GI:51874305	3617
HHV-7	U90*	IE-A transactivator; similar to HHV-5 UL123	GI:51874306	3760

Table 4 continued

HHV-7	U95	betaherpesvirus US22 gene family; possible HHV-5 TRS1	GI:51874308	2822
CeHV-2	immediate early protein ICP0	multifunctional regulatory protein	GI:56694722	2491
CeHV-2	immediate early protein ICP0	multifunctional regulatory protein	GI:56694781	2491
EHV-1	ORF 64	transcriptional activator	GI:50313305	4463
EHV-1	ORF 64	transcriptional activator	GI:50313321	4463
AtHV-3	orf 06	major ssDNA binding protein	GI:9631197	3386
AtHV-3	orf 10	similar to Raji LF1	GI:9631200	1220
AtHV-3	orf 11	similar to Raji LF2	GI:9631201	1217
AtHV-3	orf 14*	Mitogen	GI:9631202	821
AtHV-3	orf 18		GI:9631208	770
AtHV-3	orf 21*	thymidine kinase	GI:9631211	1583
AtHV-3	orf 23	similar to BTRF1	GI:9631213	767
AtHV-3	orf 33		GI:9631225	992
AtHV-3	orf 34		GI:9631226	950
AtHV-3	orf 45*		GI:9631236	782
AtHV-3	orf 48*		GI:9631239	2378
AtHV-3	orf 49*		GI:9631240	914
AtHV-3		Probable transcription activator EDRF1	GI:19343431	1343
AtHV-3	orf 55*		GI:9631246	602
AtHV-3	orf 59*		GI:9631249	1100
AtHV-3	orf 60*	small subunit of ribonucleotide reductase	GI:9631250	917
AtHV-3	orf 62		GI:9631252	992
AtHV-3	orf 64	large tegument protein	GI:9631254	7415
AtHV-3	orf 66		GI:9631256	1334
AtHV-3	orf 70*	thymidylate synthase	GI:9631261	872
AtHV-3	orf 72*	v-cyclin	GI:9631263	788
AtHV-3	orf 75		GI:9631266	3899
TuHV-1	t123*	hypothetical protein	GI:14251117	1103
HHV-8	ORF 73*	extensive acidic domains, potential leucine zipper; immediate early protein homolog	GI:18846043	3488

[a]Description as derived from GenBank annotations and other papers. *reported in previous papers or found in previous papers to have cellular homologues and suggested to be acquired from other organisms.

Discussion

In this paper, we introduced a composition-based framework for the detection of horizontal gene transfers by using SVM. This method reached a higher accuracy (over 95%) in detecting horizontally transferred genes compared to the Tsirigos and Rigoutsos's paper (2005a, b) (less than 70%) and our previous method (less than 95%) (Fu *et al.*, 2008). Using this method, 302 transferred genes were identified in 33 mammalian herpesviruses. However, in our previous paper, only 141 transferred genes were predicted (Fu *et al.*, 2008).

This paper used the SVM instead of the Mahalanobis distance, the posterior probability, or stepwise and Fischer linear discriminant analysis used in previous reports (Nakamura and Itoh, 2004; Tsirigos and Rigoutsos, 2005a; Fu *et al.*, 2008) to classify the two opposite groups. SVM is a new pattern recognition method based on statistical learning theory and has been used on a large variety of problems, including text classification (Joachims, 1998, 1999), image recognition tasks, bioinformatics and medical applications. It showed many advantages in the classification of small samples, nonlinear and multidimensional data. The algorithm has scalable memory requirements and can handle problems with many thousands of support vectors efficiently.

Although SVM has been used to detect the HGT of the herpesviruses in Tsirigos and Rigoutsos's paper (2005b), they used this algorithm to analyze only one herpesvirus – HHV-5, and reached the accuracy of less than 70%. In their analysis, the general signatures of most genes in the genome were used for discrimination: a gene with a signature outside of the general signatures of most genes in this genome would be considered as transferred gene. However, by training the learning module of SVM*light* for subsequent horizontally transferred gene detection with the conserved genes of viruses as the dataset of non-transferred genes and the conserved genes of mammals as the dataset of transferred genes, our method avoids artificial setting of a threshold for discrimination, as the previous method does, which calculates the distance between the gene and the genome. In our method, all mammalian herpesviruses were considered as a cluster to be analyzed because they have similar genome composition. This increased the amount of the analyzed data, but it also enlarged the distribution scope of the points of gene compositional signature, and probably decreased the exactness of the analysis. Fortunately, the opposite data are the host genomes, which have very different gene compositional signature, so that using all mammalian herpesviruses would not influence the exactness of the result. For the same reason, using the dinucleotides as the template to generalize the compositional signatures of genes have yielded the best discrimination results, and this procedure required to process only 4^2 data for every gene, comparing to 4^8 data for every gene in previous reports, which used 8-ker oligonucleotide template (Tsirigos and Rigoutsos, 2005a,b) and 4^3 in our previous paper, which used trinucleotide template. The result indicated that our method is perfect for the detection of herpesvirus transferred genes that could be recently acquired from the mammalian hosts (some genes transferred anciently were not detectable, the reason see below). This method can be expanded to the detection of transferred genes from hosts other than mammals just by training learning module of SVM*light* with the conserved genes from those hosts as the dataset of transferred genes.

Short sequences (<500 bp) often appear atypical for stochastic reasons and might be misidentified as having been transferred (Lawrence and Ochman, 2002). In order to overcome this problem, we avoided using short conserved genes (for example UL49A in HHV-1) as the non-transferred genes and deleted all short sequences (<500 bp) in our sequence datasets.

Although the function of most of the transferred genes detected with this method is unknown, many had been determined or predicted to be glycoproteins, membrane proteins or involved in the interaction with host, for example, dealing with the immune response, such as IL-10, cell apoptosis such as Bcl-2, and cell proliferation control, such as cyclin and mitogen, which was well consistent with other investigators' conclusion (Raftery *et al.*, 2000; Holzerlandt *et al.*, 2002; Fu *et al.*, 2008).

131 transferred genes detected by our method had been reported to be transferred genes in previous studies or to have cellular homologues and had been suggested that they might have been pilfered from their host (marked with * in Table 4). For example, the beta-1,6-N-acetyl-glucosaminyltransferase was suggested to have been acquired from an ancestor of the buffalo after the origin of the *Bovinae* (Markine-Goriaynoff *et al.*, 2003). UL122 and UL123 in HHV-5 had been predicted by Tsirigos and Rigoutsos (Tsirigos and Rigoutsos, 2005b). IL-10 in HHV-4 and CeHV-15 was believed to have a eukaryotic origin in many previous papers (Raftery *et al.*, 2000; Holzerlandt *et al.*, 2002; Hughes, 2002; Fu *et al.*, 2008). Herpesvirus homologues of cellular genes include G protein-coupled receptor (GCR) or its homologues in EHV-2, CalHV-3, CeHV-15, MuHV-4, AlHV-1 and HHV-4, Bcl-2 or its homologues in MuHV-4, CeHV-15, AlHV-1 and BoHV-4, DNA polymerase or its homologues in CeHV-15, HHV-4, AlHV-1, BoHV-4, MuHV-4 and SaHV-2, dihydrofolate reductase in SaHV-2, small subunit of ribonucleotide reductase in AlHV-1, BoHV-4, MuHV-4, SaHV-2, AtHV-3, CeHV-15 and HHV-4, thymidylate synthase in CeHV-15, BoHV-4, EHV-2, MuHV-4, SaHV-2 and AtHV-3, v-FGAM-synthetase or its homologues in HHV-4, AlHV-1, BoHV-4, MuHV-4 and SaHV-2, cyclin or its homologues in MuHV-4, SaHV-2 and AtHV-3 (Raftery *et al.*, 2000; Holzerlandt *et al.*, 2002). Apart from these predicted genes, other transferred genes detected in this study were identified as horizontally transferred genes for the first time.

As for these firstly identified transferred genes, some of them are homologues of transferred genes, which had been predicted previously. For example, the ORF19 in CalHV-3 is homologous to the BDLF2 in HHV-4, and BFRF2 in CeHV-15 is similar to that in HHV-4. The latters were predicted as transferred genes in our previous paper. The formers were identified in this research. This may be due to the high sensitivity of SVM method over others. The high sensitivity could also mark some non-transferred genes as transferred genes because of their only a little atypical composition, which may result from factors other than transfer. So we had to combine different methods to analyze the transferred genes.

Many of the transferred genes predicted in this paper were linked with some of the transferred genes identified previously. For example, BCRF2 and BWRF1 in HHV-4 are linked with BCRF1 (IL-10). BMRF1 and BMRF2 in HHV-4 are linked with BaRF1 (ribonucleotide reductase, small subunit). ORF58 in AlHV-1 is linked with ORF59 (DNA polymerase). ORF73 and A10 in AlHV-1 are linked with ORF75 (formylglycineamide-synthase) and A9 (Bcl-family of proteins). ORF10 in BoHV-4 is linked with ORF9 (DNA polymerase). ORF58 and ORF59 in BoHV-4 are linked with ORF60 (ribonucleotide reductase, small subunit). ORFE8 and ORF75 in EHV-2 are linked with ORF74 (GCR). ORF4 and ORF6 in MuHV-4 are linked with M4 (GCR homologue). M6 and ORF16b in MuHV-4 are linked with ORF21 (thymidine kinase). ORF10 in SaHV-2 is linked with ORF9 (DNA polymerase). ORF 71 in SaHV-2 is linked with ORF72 (cycline homologue). ORF10 and ORF11 in AtHV-3 are linked with ORF14. HGT always occurred in cluster, so these newly identified genes may have been horizontally transferred together with their linked genes.

Some transferred genes, which might have been transferred very early and have undergone long time of "amelioration", might already have the features of the recipient genomes (Lawrence and Ochman, 1997) and already lost their atypical characteristics. Genes like these were not detectable by this method and were considered as non-transferred genes. For example, the UL2 and UL50, which have been reported to be transferred from their host very early (Baldo and McClure, 1999; Holzerlandt *et al.*, 2002; Davison and Stow, 2005), were considered as non-transferred genes in our study.

Unusual nucleotide composition was what this method was based on. Some conserved genes were identified as transferred genes in this paper due to their atypical composition such as BcLF1 in CeHV-15 and ORF25 in AlHV-1, BoHV-4, MuHV-4 and SaHV-2, which encode the major capsid protein, UL71 in PoHV-4 and ORF55 in AlHV-1, which encode tegument protein, U53.5 in HHV-7 and ORF 17.5 in EHV-2, which encode the capsid scaffold protein. Although unusual nucleotide composition mainly resulted from the transfer events, other reasons may also cause unusual nucleotide composition. For example, BZLF and RTA transactivators as well as EBNA-1 and IE1, IE2 proteins contain anomalous clusters of charged amino acid residues (Karlin *et al.*, 1989), which would result in unusual nucleotide composition and would bias the results towards identifying these genes as transferred genes. In this case, the unusual nucleotide composition is the result of the function of the gene itself rather than recent HGT. Transfer events may sometimes not lead to an unusual composition because of the "amelioration" or the similar composition of the donor. These factors limit the development of detection methods based on composition to some extent. Our method is based on the result of unusual composition from the transfer events. Its accuracy reached the value of over 95%, though the sensitivity of detection for every herpesvirus may be different because of their different evolution rates.

When the result of this paper was compared with that of our previous paper, which used the discrimination package of SPSS to identify HGTs, it was found that there were 93 genes identified by both composition-based methods. Other genes identified by only one composition-based method may be the result of different accuracy or transferring at different time. We also compared our composition-based method with the method based on the similarity search, the earliest and the most intuitive way of identifying horizontally acquired genes. In the similarity method, gene transfers are recognized by an unusually high level of similarity among genes found in otherwise unrelated organisms. Using this method, altogether 23 groups of proteins were identified to be horizontally transferred genes (Table 5). Among them, 14 groups (marked * in Table 5) were detected by the composition-based method, other 9 groups were identified only by the similarity method. So the results of these two prediction methods overlap to some extent. This occurs because each of the methods used to detect HGT recognized different features in their target genes and are thus testing different types of hypotheses. The impact of HGT on the entire evolution of a lineage must be inferred from present-day sequences, and the different approaches used to recognize HGT must rely on specific models of sequence evolution. The different methods are based on different assumptions. The similarity-search method identifies genes, whose closest homologues are found in taxa not otherwise related to the query genome, and thus it uncovers a set of genes biased towards those that have been transferred across large phylogenetic distance, regardless of their time of arrival into a genome. On the other hand, the composition-based method examines sequence features and preferentially identifies genes that have been recently introduced into a genome from an organism having different mutational biases, regardless of phylogenetic distance. Assuming that the frequency of transfer between lineages is inversely related to their phylogenetic distance, these two methods would identify quite different sets of genes (Lawrence and Ochman, 2002). This phenomenon has been observed by Ragan (Ragan, 2001), who used the different methods to recognize significantly different subsets of genes as being subject to horizontal transfer.

Every method has its own advantage and disadvantage in identification of HGT. The composition-based method can

Table 5. Horizontally transferred genes identified by the similarity method

Virus taxon	Virus/Eukaryote function
Alpha, Beta, Gamma	Uracil DNA glycosylase
Alpha	Large tegument protein
Alpha, Gamma	Ribonucleotide reductase large subunit
Alpha, Gamma	*Ribonucleotide reductase small subunit
Alpha	Ser/Thr protein kinase
Alpha	Transactivating tegument protein
Alpha, Gamma	*DNA polymerase
Alpha, Gamma	*Thymidylate synthase
Beta	rh10 (CeHV-8); Prostaglandin synthase, cyclooxygenase-2 (other organisms)
Gamma	*FGAM
Gamma	*IL-10
Gamma	*Semaphorin 7A
Gamma	*β-1,6-actylglucosaminyltranferase
Gamma	*Complement binding protein
Gamma	*Cyclin D homologue
Gamma	CD200antigen, OX-2
Gamma	*GCR (IL-8)
Gamma, Beta	GCR (CC chemokine receptor)
Gamma	*DHFR
Gamma	*Complementary control protein
Gamma	IL-17
Gamma	*collagen
Gamma	*Latent nuclear antigen

*Identified by the SVM-based method.

escape from the complicated phylogenetic analysis, it requires a completed genome and can employ many mathematic models to design the arithmetic. But it has its limitation as illuminated above. Transferred gene is not the only reason of atypical composition. Evolution selection, mutation preference and the different orientation of transcription can influence the composition. The genes transferred from a donor having similar composition and the ameliorated genes transferred anciently can not be detected. The similarity-search method is simple and intuitive, but it also has its limitation: first, horizontally transferred gene is not the only mechanism that produces conflicts between phylogenies. Some genes might be coincidentally deleted from multiple lineages, leading to unusual distributions among extant organisms, or similarity can result from convergent evolution. Moreover, the proliferation of gene families can make the identification of orthologous sequences difficult, and rapid sequence evolution makes alignment of homologous sites equivocal. Second, the result of the similarity-search method is limited by the capacity of the search database. Third, the level of similarity is a man-made factor, if it is set too high, the sensitivity will be very low, if it is set too low, the accuracy will be also very low. In our analysis, the similarity score is 100 and this is relatively high. So using this method still led to the loss of many horizontally transferred genes, but this can be compensated by the atypical composition identification. Fourth, the similarity-search method can not decide the direction of transfer. Horizontally transferred gene may be transferred from the virus to the host, but also from the host to the virus. Fifth, the similarity usually combines the phylogenetic analysis to predict the HGT, but this deduction could be incorrect due to the low number of sequences used for analysis or the incredible phylogenetic relationship. For example, most phylogenetic trees showed that the rates of evolution of viral genes were faster than the evolution rates in the genes of other organism. This may lead to the most frequent problem, the long branch attraction. To solve this problem, more and nearer sequences should be analyzed and a more excellent method for reconstruction of phylogenetic trees should be used. So whenever possible, application of a variety of methods provides the best information about the scope of gene transfer across broad timescales.

Acknowledgments. This work was funded by the National Science Foundation of Guangdong Province under Grant number 04009785, and a grant from 211 Project of Guangdong University of Technology.

References

Alcami A, Koszinowski UH (2000): Viral mechanisms of immune evasion. Trends Microbiol. 8, 410–418. doi.org/10.1016/S0966–842X(00)01830–8

Altschul SF, Madden TL, Schaffer AJ, Zhang J, Zhang Z, Miller W, Lipman DJ (1997): Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. 25, 3389–2402. doi.org/10.1093/nar/25.17.3389

Baldo AM, McClure MA (1999): Evolution and horizontal transfer of dUTPase-encoding genes in viruses and their hosts. J. Virol. 73, 7710–7721.

Davison AJ, Stow ND (2005): New genes from old: redeployment of dUTPase by herpesviruses. J. Virol. 79, 12880–12892. doi.org/10.1128/JVI.79.20.12880–12892.2005

Ensser A, Pflanz R, Fleckenstein B (1997): Primary structure of the alcelaphine herpesvirus 1 genome. J. Virol. 71, 6517–6525.

Fu MH, Deng RQ, Wang JW, Wang XZ (2008): Detection and analysis of horizontal gene transfer in Herpesvirus. Virus Res. 131, 65–76. doi.org/10.1016/j.virusres.2007.08.009

Garcia-Vallvé S, Romeu A, Palau J (2000): Horizontal gene transfer in bacterial and archaeal complete genomes. Genome Res. 10, 1719–1725.

Garcia-Vallvé S, Guzman E, Montero MA, Romeu A (2003): HGT–DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. Nucleic Acids Res. 31, 187–189.

Holzerlandt R, Orengo C, Kellam P, Alba MM (2002): Identification of new herpesvirus gene homologs in the human genome. Genome Res. 12, 1739–1748. doi.org/10.1101/gr.334302

Hughes AL (2002): Origin and evolution of viral interleukin-10 and other DNA virus genes with vertebrate homologues. J. Mol. Evol. 54, 90–101. doi.org/10.1007/s00239–001–0021–1

Jiang H, Cho YG, Wang F (2000): Structural, functional, and genetic comparisons of Epstein-Barr virus nuclear antigen 3A, 3B, and 3C homologues encoded by the rhesus lymphocryptovirus. J. Virol. 74, 5921–5932.

Joachims T (1998): Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Proceedings of the European Conference on Machine Learning. Dortmund, pp. 137–142.

Joachims T (1999): Transductive Inference for Text Classification using Support Vector Machines. International Conference on Machine Learning (ICML). Bled, pp. 200–209.

Johannsen E, Luftig M, Chase MR, Weicksel S, Cahir-McFarland E, Illanes D, Sarracino D, Kieff E (2004): Proteins of purified Epstein-Barr virus. Proc. Natl. Acad. Sci. USA 101, 16286–16289. doi.org/10.1073/pnas.0407320101

Karlin S, Blaisdell BE, Mocarski ES, Brendel V (1989): A method to identify distinctive charge configurations in protein sequences, with application to human herpesvirus polypeptides. J. Mol. Biol. 205, 165–177. doi.org/10.1016/0022–2836(89)90373–2

Lawrence JG, Ochman H (1997): Amelioration of bacterial genomes: rates of change and exchange. J. Mol. Evol. 44, 383–397. doi.org/10.1007/PL00006158

Lawrence JG, Ochman H (1998): Molecular archaeology of the Escherichia coli genome. Proc. Natl. Acad. Sci. USA 95, 9413–9417. doi.org/10.1073/pnas.95.16.9413

Lawrence JG, Ochman H (2002): Reconciling the many faces of lateral gene transfer. Trends Microbiol. 10, 1–4. doi.org/10.1016/S0966–842X(01)02282–X

Markine-Goriaynoff N, Georgin JP, Golta M, Zimmermann W, Broll H, Wamwayi HM, Pastoret PP, Sharp PM, Vanderplasschen A (2003): The core 2β-1,6-N-acetylglucosaminyltransferase-mucin encoded by bovine herpesvirus 4 was acquired from an ancestor of the African Buffalo. J. Virol. 77, 1784–1792. doi.org/10.1128/JVI.77.3.1784–1792.2003

McFadden G, Murphy PM (2000): Host-related immunomodulators encoded by poxviruses and herpesviruses. Curr. Opin. Microbiol. 3, 371–378. doi.org/10.1016/S1369–5274(00)00107–7

Moore PS, Bossoff C, Weiss RA, Chang Y (1996): Molecular mimicry of human cytokine and cytokine response pathway genes by KSHV. Science 274, 1739–1744. doi.org/10.1126/science.274.5293.1739

Mrazek J, Karlin S (1999): Detecting alien genes in bacterial genomes. Ann. NY. Acad. Sci. 870, 314–329. doi.org/10.1111/j.1749–6632.1999.tb08893.x

Nakamura Y, Itoh T, Matsuda H (2004): Gojobori T. Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nature Gentics 36, 760–766. doi.org/10.1038/ng1381

Ochman H, Lawence JG, Groisman EA (2000): Lateral gene transfer and the nature of bacterial innovation. Nature 405, 299–304. doi.org/10.1038/35012500

Paulsen SJ, Rosenkilde MM, Eugen-Olsen J, Kledal TN (2005): Epstein-Barr virus-encoded BILF1 is a constitutively active G protein-coupled receptor. J. Virol. 79, 536–546. doi.org/10.1128/JVI.79.1.536–546.2005

Raftery M, Muller A, Schonrich G (2000): Herpesvirus homologues of cellular genes. Virus Genes 21, 65–75. doi.org/10.1023/A:1008184330127

Ragan MA (2001): On surrogate methods for detecting lateral gene transfer. FEMS Microbiol. Lett. 210, 187–191. doi.org/10.1111/j.1574–6968.2001.tb10755.x

Rivailler P, Jiang H, Cho Y, Quink C, Wang F (2002): Complete nucleotide sequence of the rhesus lymphocryptovirus: genetic validation for an Epstein-Barr virus animal mode. J. Virol. 76, 421–426. doi.org/10.1128/JVI.76.1.421–426.2002

Shackelton LA, Holmes EC (2004): The evolution of large DNA viruses: combining genomic information of viruses and their hosts. Trends Microbiol. 12, 458–465. doi.org/10.1016/j.tim.2004.08.005

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA (2003): The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4, 41–44. doi.org/10.1186/1471–2105–4–41

Tatusov RL, Koonin EV, Lipman DJ (1997): A genomic perspective on protein families. Science 278, 631–637. doi.org/10.1126/science.278.5338.631

Tsirigos A, Rigoutsos I (2005a): A new computational method for the detection of horizontal gene transfer events. Nucleic Acids Res. 33, 922–933. doi.org/10.1093/nar/gki187

Tsirigos A, Rigoutsos I (2005b): A sensitive, support-vector-machine method for the detection of horizontal gene transfers in viral, archaeal and bacterial genomes. Nucleic Acids Res. 33, 3699–3707. doi.org/10.1093/nar/gki660

Willms AR, Rouqhan PD, HeineMann JA (2006): Static recipient cells as reservoirs of antibiotic resistance during antibiotic therapy. Theor. Popul. Biol. 70, 436–451. doi.org/10.1016/j.tpb.2006.04.001