Classifying Nucleotide Sequences and their Positions of Influenza A Viruses through Several Kernels*

Issei Hamada[1], Takaharu Shimada[1†], Daiki Nakata[1], Kouichi Hirata[1] and Tetsuji Kuboyama[2]
[1]*Kyushu Institute of Technology, Kawazu 680-4, Iizuka 820-8502, Japan*
[2]*Gakushuin University, Mejiro 1-5-1, Toshima, Tokyo 171-8588, Japan*

Keywords: Kernels, Nucleotide Sequences, Positions in Nucleotide Sequences, Phylogenetic Trees.

Abstract: In this paper, we classify nucleotide sequences and their positions of influenza A viruses by using both *nucleotide sequence kernels* and *phylogenetic tree kernels*. In the nucleotide sequence kernel, we regard a nucleotide sequence as a vector, a multiset and a string. In the phylogenetic tree kernel, we use a *relabeled phylogenetic tree* obtained by replacing the labels of leaves that are indices of nucleotide sequences in the reconstructed phylogenetic tree from a set of nucleotide sequences with the nucleotides at a fixed position and *trimmed phylogenetic trees* obtained by trimming the branches in the relabeled phylogenetic tree with same leaves as possible. Then, we observe which of kernels are effective the classification of nucleotide sequences as analyzing pandemic occurrences and regions and the classification of positions in nucleotide sequences as analyzing positions in packaging signals.

1 INTRODUCTION

In this paper, we classify nucleotide sequences and their positions of influenza A viruses by using *nucleotide sequence kernels* and *phylogenetic tree kernels* through LIBSVM (Chang and Lin, 2013).

In the nucleotide sequence kernels, we use a *naïve kernel*, a *multiset kernel* (Gärtner, 2008) and a *spectrum string kernel* (Leslie et al., 2002) by regarding a nucleotide sequence as a vector, a multiset and a string, respectively.

On the other hand, in the phylogenetic tree kernels, we prepare *relabeled phylogenetic trees* obtained by replacing the labels of leaves that are indices of nucleotide sequences in the reconstructed phylogenetic tree from a set of nucleotide sequences with the nucleotides at a fixed position, and *trimmed phylogenetic trees* obtained by trimming the branches in the relabeled phylogenetic tree with same leaves as possible. Then, we use an *agreement subtree mapping kernel* (Hamada et al., 2013) and a *leaf-path kernel* to classify relabeled or trimmed phylogenetic trees.

As the target of classification of nucleotide sequences, we classify nucleotide sequences of pandemic viruses from ones of non-pandemic viruses, called *pandemic classification*, and nucleotide sequences at one region from ones at other regions, called *regional analysis*, for influenza A (H1N1) viruses as similar as (Hamada et al., 2013; Makino et al., 2012b; Shimada et al., 2013). As the target of classification of positions in nucleotide sequences, we classify positions in packaging signals from ones not in packaging signals, called *packaging signal analysis* for influenza A (H3N2) viruses as similar as (Makino et al., 2012a; Shimada et al., 2012).

Hence, we observe that both the nucleotide sequence kernels and the phylogenetic tree kernels are effective to the pandemic classification. Also the nucleotide sequence kernels and the leaf-path kernel are effective to the packaging signal analysis. Furthermore, the phylogenetic tree kernels but none of nucleotide sequence kernels are effective to the regional analysis.

2 NUCLEOTIDE SEQUENCE KERNELS

Let Σ be $\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}$ with an alphabetical order $\preceq$. Throughout of this paper, we assume that a nucleotide sequence is a sequence on Σ and every sequence in a

*This work is partially supported by Grant-in-Aid for Scientific Research 24240021, 24300060, 25540137, 26280085, 26280090 and 26370281 from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

†Current Affilication: Mazda Motor Corporation.

Hamada I., Shimada T., Nakata D., Hirata K. and Kuboyama T..
Classifying Nucleotide Sequences and their Positions of Influenza A Viruses through Several Kernels.
DOI: 10.5220/0005251103420347
In *Proceedings of the International Conference on Pattern Recognition Applications and Methods* (ICPRAM-2015), pages 342-347
ISBN: 978-989-758-076-5

Copyright © 2015 SCITEPRESS (Science and Technology Publications, Lda.)

set of nucleotide sequences has the same length.

First, we regard a nucleotide sequence as a vector on Σ. For $x,y \in \Sigma$, we define $\delta_1(x,y) = 1$ if $x = y$; 0 otherwise. Also we define $\delta_2(x,y) = 1$ if $x = y$; $1/2$ if $(x,y) = (\mathtt{A},\mathtt{T}),(\mathtt{T},\mathtt{A}),(\mathtt{C},\mathtt{G}),(\mathtt{G},\mathtt{C})$ (that is, base pairs are weighted); 0 otherwise. Then, we define a *naïve kernel* K_j $(j = 1,2)$ for two vectors $X = (x_1,\ldots,x_n)$ and $Y = (y_1,\ldots,y_n)$ $(x_i,y_i \in \Sigma)$ on Σ as follows.

$$K_j(X,Y) = \frac{1}{n}\sum_{i=1}^{n}\delta_j(x_i,y_i).$$

Next, we regard a nucleotide sequence as a *multiset*. We call $X \subseteq \Sigma \times \mathbf{N}$ a *multiset* on Σ. For a multiset X, let $\Gamma_X(x)$ denote an n such that $(x,n) \in X$. Then, we define a *multiset product kernel* $K_\times$ and a *multiset intersection kernel* $K_\cap$ for two multisets X and Y on Σ as follows.

$$\begin{aligned} K_\times(X,Y) &= \sum_{a\in\Sigma}\Gamma_X(a)\cdot\Gamma_Y(a), \\ K_\cap(X,Y) &= \sum_{a\in\Sigma}\min\{\Gamma_X(a),\Gamma_Y(a)\}. \end{aligned}$$

Finally, we regard a nucleotide sequence as a string on Σ. For a string $X \in \Sigma^*$ and a substring $s \in \Sigma^*$ of X, let $\Gamma_X(s)$ be the number of occurrences of s in X. Also, for $k \in \mathbf{N}$, let Σ^k be $\{s \in \Sigma^* \mid |s| = k\}$. Then, we define a *spectrum string kernel* K_S^k for two strings X and Y on Σ as follows.

$$K_S^k(X,Y) = \sum_{s\in\Sigma^k}\Gamma_X(s)\cdot\Gamma_X(s).$$

3 PHYLOGENETIC TREE KERNELS

A *tree* is a connected graph without cycles. For a tree $T = (V,E)$, we denote V by $V(T)$ and $v \in V(T)$ by $v \in T$. A *rooted tree* is a tree with one node r chosen as its *root*.

For each node v in a rooted tree with the root r, let $UP_r(v)$ be the unique path from r to v. The *parent* of $v(\neq r)$, which we denote by $par(v)$, is its adjacent node on $UP_r(v)$ and the *ancestors* of $v(\neq r)$ are the nodes on $UP_r(v) - \{v\}$. We say that u is a *child* of v if v is the parent of u, and u is a *descendant* of v if v is an ancestor of u.

In this paper, we use the ancestor orders $<$ and $\leq$, that is, $u < v$ if v is an ancestor of u and $u \leq v$ if $u < v$ or $u = v$. We say that w is the *least common ancestor* of u and v, denoted by $u \sqcup v$, if $u \leq w$, $v \leq w$, and there exists no w' such that $w' < w$, $u \leq w'$ and $v \leq w'$.

Two nodes with the common parent are called *siblings*. A *leaf* is a node having no children. We denote the set of leaves of a rooted tree T by $lv(T)$. For nodes $v, w \in V$, we denote a *path* between v and w by $p(v,w)$. Also we denote the number of edges in a path $p(v,w)$ by $ne(v,w)$. It is obvious that $ne(v,v) = 0$.

A rooted tree is *unordered* if an order between siblings is ignored. A rooted tree is *leaf-labeled* if just leaves are labeled by some symbols drawn from Σ and *full binary* if every internal node has just two children. We denote the label of a leaf v in Σ by $l(v)$. We call a rooted unordered leaf-labeled full binary tree a *phylogenetic tree*. As a reconstruction of a phylogenetic tree from a set of nucleotide sequences, we adopt a *neighbor joining method* (*cf.*, (Durbin et al., 1998; Sung, 2009)) based on the Hamming distance between nucleotide sequences.

Let S be a set of nucleotide sequences with length n and T a phylogenetic tree reconstructed from S. Then, we can obtain n phylogenetic trees by relabeling an index of S assigned to the leaves in T with the i-th nucleotide in S $(1 \leq i \leq n)$, which we call a *relabeled phylogenetic tree* at the position i. Furthermore, we call the phylogenetic tree obtained by applying the *label-based closest-neighbor trimming method* (Makino et al., 2012b; Makino et al., 2012a) to the relabeled phylogenetic tree at the position i the *trimmed phylogenetic tree* at the position i.

In the remainder of this section, we introduce an agreement subtree mapping kernel and a leaf-path kernel as phylogenetic tree kernels.

Let T_1 and T_2 be phylogenetic trees. Then, we say that $M \subseteq V(T_1) \times V(T_2)$ is a *mapping* between T_1 and T_2 if M satisfies the following conditions.

1. $\forall (v_1,w_1),(v_2,w_2) \in M \Big(v_1 = v_2 \iff w_1 = w_2\Big)$.
2. $\forall (v_1,w_1),(v_2,w_2) \in M \Big(v_1 \leq v_2 \iff w_1 \leq w_2\Big)$.

Let T_1 and T_2 be phylogenetic trees and M a mapping between T_1 and T_2. Also let M^{lv} be $M \cap (lv(T_1) \times lv(T_2))$. Then, we say that M is an *agreement subtree mapping* (Hamada et al., 2013) if M satisfies the following conditions.

1. $\forall (v,w) \in M \Big(v \in lv(T_1) \iff w \in lv(T_2)\Big)$.
2. $\forall (v,w) \in M^{lv} \Big(l(v) = l(w)\Big)$.
3. $\forall (v_1,w_1),(v_2,w_2) \in M^{lv} \Big((v_1 \sqcup v_2, w_1 \sqcup w_2) \in M\Big)$.
4. $\forall (v,w) \in M - M^{lv}\ \exists (v_1,w_1),(v_2,w_2) \in M^{lv}$ $\Big((v = v_1 \sqcup v_2) \wedge (w = w_1 \sqcup w_2)\Big)$.

Definition 1. Let T_1 and T_2 be phylogenetic trees. Then, an *agreement subtree mapping kernel* is the number of all of the agreement subtree mappings between T_1 and T_2 and denote it by $K_{\mathrm{AM}}(T_1,T_2)$.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

For example, consider the tree T illustrated in Figure 1 (left). Then, Figure 1 (right) illustrates all of the agreement subtree mappings between T and T. Hence, it holds that $K_{\text{AM}}(T,T) = 6$.

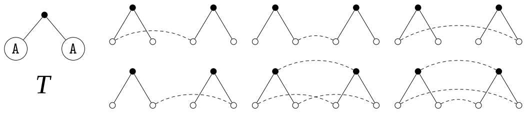


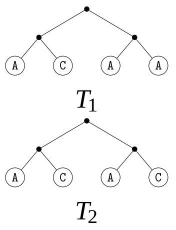
Figure 1: The tree T (left) and all of the agreement subtree mappings between T and T (right).

For a phylogenetic tree T and $v,w \in lv(T)$, we denote the frequency of a path $p(v,w)$ such that $l(v) = a$, $l(w) = b$ and $ne(v,w) = k$ by $f_T(a,b,k)$.

Definition 2. Let T_1 and T_2 be phylogenetic trees labeled by Σ. Then, the *leaf-path kernel* $K_{LP}(T_1,T_2)$ between T_1 and T_2 is defined as follows, where $\Delta = 2 \cdot \max\{dep(T_1), dep(T_2)\}$.

$$K_{\text{LP}}(T_1,T_2) = \sum_{a\in\Sigma} \sum_{b\in\Sigma, a\preceq b} \sum_{k=0}^{\Delta} f_{T_1}(a,b,k) \cdot f_{T_2}(a,b,k).$$

For example, consider the trees T_1 and T_2 in Figure 2 (upper). Then, we obtain $f_{T_1}(a,b,k)$ and $f_{T_2}(a,b,k)$ as Figure 2 (lower). Hence, it holds that $K_{\text{LP}}(T_1,T_2) = 16$.



a b k	f_{T_1}	f_{T_2}
A A 0	3	2
A A 2	1	0
A A 4	2	1
A C 2	1	2
A C 4	2	2
C C 0	1	2
C C 4	0	1

Figure 2: Trees T_1 and T_2 (left) and $f_{T_1}(a,b,k)$ and $f_{T_2}(a,b,k)$ (right).

We denote an agreement subtree mapping kernel (*resp.*, a leaf-path kernel) for trimmed and relabeled phylogenetic trees by K^t_{AM} and K^r_{AM} (*resp.*, K^t_{LP} and K^r_{LP}), where we use K^r_{AM} just in Table 2.

4 CLASSIFICATION OF NUCLEOTIDE SEQUENCES

In the classification of nucleotide sequences, we divide a set of nucleotide sequences into positive and negative examples. Then, in the phylogenetic tree kernels, we use two different phylogenetic trees reconstructed from positive and negative examples, respectively. Hence, the number of relabeled and trimmed phylogenetic trees obtained from positive examples is same as one from negative examples, which is the length of nucleotide sequences. On the other hand, the number of leaves in a relabeled phylogenetic tree obtained from positive examples is different from one from negative examples, which is the number of nucleotide sequences.

4.1 Pandemic classification

In pandemic classification, we use 3670 nucleotide sequences at 2008 and 2009 provided from NCBI (Bao et al., 2008). The length of nucleotide sequences is 895, the number of nucleotide sequences in non-pandemic viruses occurring at 2008 is 326 and one in pandemic viruses occurring at 2009 is 3344.

Table 1 illustrates the F-value and the AUC of 5-fold cross validation classifying nucleotide sequences in non-pandemic viruses from ones in pandemic viruses by using all the kernels through LIBSVM (Chang and Lin, 2013).

Table 1: The classification of nucleotide sequences in non-pandemic viruses from ones in pandemic viruses.

	K_1	K_2	$K_\times$	$K_\cap$	K_S^1	K_S^2	K_S^3	K_S^4	K_S^5	K^t_{AM}	K^t_{LP}	K^r_{LP}
F-value	1	1	0.999	0.999	1	1	1	1	1	0.911	0.915	1
AUC	1	1	0.999	0.999	1	1	1	1	1	0.951	0.866	1

In order to avoid the bias of the number of examples, Table 2 illustrates the F-value and the AUC of 5-fold cross validation after randomly selecting 200 nucleotide sequences from 2008 and 2009, respectively.

Table 2: The classification of randomly selected 400 nucleotide sequences in non-pandemic viruses from ones in pandemic viruses.

	K_1	K_2	$K_\times$	$K_\cap$	K_S^1	K_S^2	K_S^3	K_S^4	K_S^5	K^t_{AM}	K^r_{AM}	K^t_{LP}	K^r_{LP}
F-value	1	1	0.995	0.997	1	1	1	1	1	0.975	0.975	1	1
AUC	1	1	0.998	0.995	1	1	1	1	1	0.992	0.998	1	1

Table 1 and 2 shows that, in the pandemic classification, all of the nucleotide sequence and the phylogenetic tree kernels succeed to classify well.

4.2 Regional Analysis

In regional analysis as an extension of the experimental result of (Hamada et al., 2013), we divide 3670 nucleotide sequences at 2008 and 2009 into seven regions as Africa (AF), Asia (AS), Europe (EU), Middle East (ME), North America (NA), Oceania (OC) and South America (SA). Table 3 illustrates the number of nucleotide sequences (#NS) and the number

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

of phylogenetic trees (#PT) obtained by removing the positions with the same nucleotide in seven regions.

Table 3: The number of nucleotide sequences (#NS) and the number of phylogenetic trees (#PT) in seven regions.

	AF	AS	EU	ME	NA	OC	SA	total
#NS	61	949	965	71	1403	47	174	3670
%	1.66	25.86	26.29	1.93	38.23	1.28	4.74	
#PT	289	593	487	311	538	290	344	2852
%	10.13	20.79	17.08	10.90	18.86	10.17	12.06	

Table 4 illustrates the F-value and the AUC of 5-fold cross validation classifying nucleotide sequences in one region given at the first line from nucleotide sequences in the other regions by using all the kernels.

Table 4: The classification of nucleotide sequences in one region given at the first line from ones in the other regions.

		AF	AS	EU	ME	NA	OC	SA
K_1	F-value	0	0.029	0	0	0	0	0
	AUC	0.622	0.690	0.657	0.636	0.662	0.743	0.645
K_2	F-value	0	0.012	0	0	0	0	0
	AUC	0.628	0.689	0.650	0.559	0.662	0.745	0.646
$K_\times$	F-value	0	0	0	0	0	0	0
	AUC	0.437	0.501	0.541	0.437	0.544	0.549	0.470
$K_\cap$	F-value	0	0.127	0.094	0	0.257	0	0
	AUC	0.445	0.550	0.616	0.499	0.593	0.637	0.562
K_S^1	F-value	0	0	0	0	0	0	0
	AUC	0.516	0.519	0.498	0.537	0.542	0.516	0.463
K_S^2	F-value	0	0.022	0	0	0.478	0	0
	AUC	0.495	0.612	0.596	0.452	0.666	0.531	0.660
K_S^3	F-value	0	0.388	0.351	0	0.480	0	0.127
	AUC	0.713	0.708	0.708	0.550	0.720	0.624	0.825
K_S^4	F-value	0.382	0.534	0.507	0	0.546	0.155	0.375
	AUC	0.713	0.708	0.708	0.550	0.720	0.624	0.825
K_S^5	F-value	0.361	0.600	0.544	0.152	0.593	0.282	0.361
	AUC	0.793	0.786	0.759	0.653	0.763	0.763	0.934
K_{AM}^t	F-value	0.911	0.766	0.929	0.031	0.830	0.300	0.753
	AUC	0.947	0.898	0.978	0.814	0.955	0.933	0.919
K_{LP}^t	F-value	0.873	0.802	0.962	0.853	0.637	0.881	0.837
	AUC	0,988	0.918	0.995	0.975	0.805	0.983	0.975
K_{LP}^r	F-value	1	1	1	0.998	1	1	1
	AUC	1	1	1	0.996	1	1	1

Table 4 shows that, in regional analysis, while the nucleotide sequence kernels fail to classify, the phylogenetic tree kernels succeed to classify well, except K_{AM}^t for the regions of ME and OC.

In particular, for the spectrum string kernel K_S^k, the larger value k tends to give the better performance except the regions of AF and SA; in their regions, the F-value of K_S^4 is larger than the F-value of K_S^5. Then, even if we give the value of k larger than 5, K_S^k may not give better performance in regional analysis.

Next, in order to avoid the bias of the number of examples, we apply regional analysis for every pair of regions. Table 5 illustrates the F-value and the AUC of 5-fold cross validation classifying nucleotide sequences in one region as positive examples from nucleotide sequences in another region as negative examples by using the phylogenetic tree kernels K_{AM}^t, K_{LP}^t and K_{LP}^r, respectively.

Table 5: The classification of nucleotide sequences in one region from ones in another region.

			AS	EU	ME	NA	OC	SA
AF	K_{AM}^t	F-value	0.967	1	0.940	0.989	0.949	0.994
		AUC	0.991	1	0.940	0.989	0.949	0.994
	K_{LP}^t	F-value	0.944	1	0.914	0.975	0.956	0.993
		AUC	0.985	1	0.925	0.995	0.967	0.999
	K_{LP}^r	F-value, AUC	1	1	1	1	1	1
AS	K_{AM}^t	F-value		0.963	0.982	0.914	0.987	0.885
		AUC		0.990	0.991	0.945	0.994	0.871
	K_{LP}^t	F-value		0.996	0.975	0.865	0.984	0.910
		AUC		0.999	0.984	0.921	0.994	0.936
	K_{LP}^r	F-value, AUC		1	1	1	1	1
EU	K_{AM}^t	F-value			0.998	0.944	0.998	0.989
		AUC			1	0.971	1	0.999
	K_{LP}^t	F-value			1	0.960	1	0.993
		AUC			1	0.988	1	0.999
	K_{LP}^r	F-value, AUC			1	1	1	1
ME	K_{AM}^t	F-value				0.980	0.756	0.998
		AUC				0.997	0.771	0.999
	K_{LP}^t	F-value				0.977	0.920	0.998
		AUC				0.991	0.934	0.999
	K_{LP}^r	F-value, AUC				1	1	1
NA	K_{AM}^t	F-value					0.996	0.937
		AUC					0.999	0.969
	K_{LP}^t	F-value					0.992	0.932
		AUC					0.997	0.954
	K_{LP}^r	F-value, AUC					1	1
OC	K_{AM}^t	F-value						1
		AUC						1
	K_{LP}^t	F-value						0.998
		AUC						1
	K_{LP}^r	F-value, AUC						1

Note that Table 3 shows that the difference between the number of phylogenetic trees in AF and OC is 1, one in OC and ME is 21, and one in ME and SA is 33. Even such regions, K_{AM}^t, K_{LP}^t and K_{LP}^r succeed to classify except for K_{AM}^t for the regions of ME and OC. In particular, for K_{AM}^t and K_{LP}^t the F-value and the AUC in Table 5 are larger than ones in Table 4. Furthermore, K_{LP}^r succeeds to classify completely all regions with the F-value and the AUC of 1.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

5 CLASSIFICATION OF POSITIONS IN NUCLEOTIDE SEQUENCES

In the classification of positions in nucleotide sequences, we divide positions into *positive positions* in target positions and *negative positions* not in target positions for a set of nucleotide sequences. Then, in the phylogenetic tree kernels, we use two different phylogenetic trees reconstructed from a set of nucleotide sequences at positive positions and one at negative positions, respectively. Hence, the number of leaves in a relabeled phylogenetic tree obtained from positive positions is same as one from negative positions, which is the number of nucleotide sequences. On the other hand, the number of relabeled and trimmed phylogenetic trees obtained from positive positions is different from one from negative positions, which is the length of nucleotide sequences.

5.1 Packaging Signal

The negative-sense RNA genome of the influenza A virus is composed of eight different segments, that is, PB2, PB1, PA, HA, NP, NA, MP and NS. Since influenza virions do not typically package more than eight segments, the virus has evolved a selective *packaging mechanism* which ensures that virions incorporate one copy of each of the eight segments. A *packaging signal* is a nucleotide to cause such a selective packaging mechanism (Hutchinson et al., 2010).

Through *reverse genetics*, segment-specific packaging signals have been found in unique regions adjacent to the panhandle of each segment. Table 6 represents the positions as packaging signals obtained by reverse genetics summarized by (Hutchinson et al., 2010) in Virology. Here, the column "NCBI" denotes the corresponding positions in nucleotide sequences of segments in influenza A (H3N2) viruses provided from NCBI (Bao et al., 2008). Also the column $(+)$ (*resp.*, $(-)$) denotes the total number of positive (*resp.*, negative) positions.

5.2 Packaging Signal Analysis

In packaging signal analysis, we use 1560 nucleotide sequences of influenza A (H3N2) viruses. Then, Table 7 illustrates the F-value and the AUC of 5-fold cross validation classifying positive positions from negative positions by using the nucleotide sequence and the phylogenetic tree kernels through LIBSVM (Chang and Lin, 2013). Here, we can obtain no value of K^t_{AM} for the NS segment.

Table 6: The positions in packaging signals of RNA segments (Hutchinson et al., 2010).

RNA	length	NCBI	(+)	(−)
PB2	2341	35–114, 2209–2304	174	2167
PB1	2341	38–163, 2197–2299	227	2114
PA	2233	38–124, 691–731, 742–767, 2094–2156, 2169–2176	220	2013
HA	1778	38–125, 1659–1671	99	1679
NP	1565	46–165, 1482–1526	163	1402
NA	1413	35–185, 1211–1413	352	1061
MP	1027	ε	–	–
NS	890	36–56	20	870

Table 7: The classification of positive positions from negative positions.

		PB2	PB1	PA	HA	NP	NA	NS
$K_1, K_2, K_\times,$	F-value	1	1	1	1	1	1	1
$K_\cap, K^k_S, K^r_{\mathrm{LP}}$	AUC	1	1	1	1	1	1	1
K^t_{AM}	F-value	0.925	0.559	0.797	0.308	0.942	0.867	–
	AUC	0.999	0.913	0.989	0.781	1	0.967	–
K^t_{LP}	F-value	1	0.649	0.921	0.414	1	0.951	1
	AUC	1	0.923	0.966	0.859	1	0.988	1

Next, in order to avoid the bias of the number of positions and the positions with the same nucleotide, for every RNA segment, we remove the positions in positive and negative positions where nucleotide is same. As a result, the number of positive positions of PB2 decreases from 174 to 150; PB1 from 227 to 113; PA from 220 to 87; HA from 99 to 77; NP from 163 to 64; NA from 352 to 205; NS from 20 to 11.

Table 8 illustrates the F-value and the AUC of 5-fold cross validation classifying positive positions after the removal of positions from randomly selected negative positions with the same number of positive positions by using the nucleotide sequence kernels except K^k_S and the phylogenetic tree kernels.

Hence, Table 7 and 8 show that K_1, K_2, $K_\times$, $K_\cap$ and K^r_{LP} succeed to classify the positions in packaging signals from ones not in packaging signals. In particular, the F-value of K^r_{LP} for the NS segment is smaller than the F-values for other segments. On the other hand, K^t_{AM} and K^t_{LP} do not classify well segments PA and NA and segments PA, HA and NS, respectively.

6 CONCLUSION

In this paper, we have classified nucleotide sequences

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

Table 8: The classification of positive positions from randomly selected negative positions.

		PB2	PB1	PA	HA	NP	NA	NS
K_1	F-value	0.999	1	1	1	1	0.999	1
	AUC	1	1	1	1	1	1	1
K_2	F-value	0.999	1	1	1	1	0.999	1
	AUC	1	1	1	1	1	1	1
$K_\times$	F-value	1	1	0.999	1	0.998	1	0.994
	AUC	1	1	1	1	0.999	1	0.995
$K_\cap$	F-value	1	1	0.999	1	0.997	1	0.995
	AUC	1	1	1	1	0.999	1	0.996
K^t_{AM}	F-value	0.909	0.935	0.721	0.959	0.916	0.825	–
	AUC	0.981	0.961	0.745	0.984	0.988	0.918	–
K^t_{LP}	F-value	0.966	0.912	0.818	0.603	0.984	0.944	0.521
	AUC	0.986	0.933	0.856	0.585	0.999	0.952	0.330
K^r_{LP}	F-value	1	1	1	1	1	1	0.916
	AUC	1	1	1	1	1	1	0.966

and positions in them of influenza A viruses by using the phylogenetic tree and the nucleotide sequence kernels. Then, we have observed that both the nucleotide sequence kernels and the phylogenetic tree kernels are effective to the pandemic classification. Also the nucleotide sequence kernels and the leaf-path kernel are effective to the packaging signal analysis. Furthermore, the phylogenetic tree kernels and none of nucleotide sequence kernels are effective to the regional analysis.

In the case that the phylogenetic tree kernels succeed to classify, two different phylogenetic trees reconstructed from positive and negative examples or positions work well as background knowledge in our classification. This is typical for regional analysis which the nucleotide sequence kernels fail to classify.

It is a future work to apply the regional analysis to influenza A (H3N2) viruses and the analysis of positions in packaging signals to influenza A (H1N1) viruses. It is also an important future work to compare the correlated mutations (Shimada et al., 2012) with our results and to analyze our results from the viewpoints of Virology. Furthermore, it is a future work to analyze, classify and evaluate another nucleotide sequences by using the phylogenetic tree kernels and the nucleotide sequence kernels.

REFERENCES

Bao, Y., Bolotov, P., Dernovoy, D., Kiryutin, B., Zaslavsky, L., Tatusova, T., Ostell, J., and Lipman, D. (2008). The influenza virus resource at the National Center for Biotechnology Information. *J. Virol.*, 82:596–601. Also available at: http://www.ncbi.nlm.gov/genomes/FLU/.

Chang, C.-C. and Lin, C.-J. (2013). *LIBSVM – A library for support vector machine (version 3.17)*. Available at http://www.csie.ntu.edu.tw/˜cjlin/libsvm.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press.

Gärtner, T. (2008). *Kernels for structured data*. World Scientific.

Hamada, I., Shimada, T., Hirata, K., and Kuboyama, T. (2013). Agreement subtree mapping kernel for phylogenetic trees. In *Proc. DDS 13*, pages 1–8.

Hutchinson, E. C., von Kirchbach, J. C., Gog, J. R., and Digard, P. (2010). Genome packaging in influenza A virus. *J. Gen. Virol.*, 91:313–328.

Leslie, C. S., Eskin, E., and Noble, W. S. (2002). The spectrum kernel: A string kernel for svm protein classification. In *Proc. PSB 2002*, pages 566–575.

Makino, S., Shimada, T., Hirata, K., Yonezawa, K., and Ito, K. (2012a). A trim distance between positions as packaging signals in H3N2 influenza viruses. In *Proc. SCIS-ISIS 2012*, pages 1702–1707.

Makino, S., Shimada, T., Hirata, K., Yonezawa, K., and Ito, K. (2012b). A trim distance between positions in nucleotide sequences. In *Proc. DS 2012 (LNAI 2569)*, pages 81–94.

Shimada, T., Hamada, I., Hirata, K., Kuboyama, T., Yonezawa, K., and Ito, K. (2013). Clustering of positions in nucleotide sequences by trim distance. In *Proc. IIAI AAI 2013*, pages 129–134.

Shimada, T., Hazemoto, T., Makino, S., Hirata, K., and Ito, K. (2012). Finding correlated mutations among rna segments in H3N2 influenza viruses. In *Proc. SCIS-ISIS 2012*, pages 1696–1705.

Sung, W.-K. (2009). *Algorithms in bioinformatics: A practical introduction*. Chapman & Hall/CRC.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS