人工知能学会研究会資料
SIG-FPAI-B401-01

さまざまなカーネルによる A 型インフルエンザウイルスの塩基配列解析

Analyzing Nucleotide Sequences of Influenza A Viruses through Various Kernels

濵田 一青 [1*]　島田 昂治 [1]　中田 大貴 [1]　平田 耕一 [2]
Issei Hamada[1]　Takaharu Shimada[1]　Daiki nakata[1]　Kouichi Hirata[3]

[1] 九州工業大学大学院情報工学府
[2] 九州工業大学情報工学研究院
[1] Graduate School of Computer Science and Systems Engineering
[2]Department of Artificial Intelligence
Kyushu Institute of Technology

Abstract: In this paper, we classify nucleotide sequences of influenza A viruses by using various kernels. Our kernels mainly consist of *nucleotide sequence kernels* by regarding nucleotide sequences as vectors, multisets and strings and *phylogenetic tree kernels* applied to phylogenetic trees reconstructed from a set of nucleotide sequences. Then, we evaluate that the phylogenetic tree kernels are effective to the pandemic classification and the regional analysis, while the nucleotide sequence kernels are effective to the pandemic classification and the analysis of positions in packaging signals.

1　はじめに

A 型インフルエンザウイルスの遺伝子は突然変異を起こしやすく, 表面粒子の抗原性が異なる子孫を生み出し, 流行を繰り返す. 例えば, A 型インフルエンザウイルスは, 1918 年に A 型 H1N1 亜型によるスペインかぜのパンデミックを引き起こし, その後, 1957 年に A 型 H2N2 亜型によるアジアかぜ, 1968 年に A 型 H3N2 亜型による香港かぜのパンデミックを引き起こした. また, 2009 年は, A 型 H1N1 亜型による新型インフルエンザのパンデミックも引き起こしている. そのため, A 型インフルエンザウイルスの塩基配列を解析し, 流行を予測することは, 医学的に重要な課題のひとつである.

そこで, 本論文では, インフルエンザウイルスを分類するために, **塩基配列カーネル** (nucleotide sequence kernel) と**進化系統樹カーネル** (phylogenetic tree kernel) の二種類のカーネルを導入する.

塩基配列カーネルとは, 内積の対象として, 直接インフルエンザウイルスの塩基配列を使用するカーネルである. 本論文では, 塩基配列を配列とみなして各位置を比較する**配列カーネル** (array kernel), 塩基配列を重集合とみなして重集合の直積または共通部分の要素数を比較する**重集合カーネル** (multiset kernel) [10], 塩基配列を文字列としてみなして 2 つの文字列に共通に出現する固定長の連続部分文字列の頻度を比較する**文字列カーネル** (spectrum string kernel) [5] を利用する.

進化系統樹カーネルとは, インフルエンザウイルスの塩基配列から推定された進化系統樹の葉を, 塩基配列のある位置の塩基と置換することで, 塩基配列長と同数得られる進化系統樹を類別するカーネルである. 本論文では, 木間の**合致部分木マッピング**の総数を数え上げる**合致部分木マッピングカーネル** (agreement subtree mapping kernel) [3], および, 両方の木に共通に出現する葉間パスの頻度を比較する**葉間パスカーネル** (leaf-path kernel) [6] を導入する. なお, 進化系統樹としては, 塩基配列から**ラベルに基づく近隣剪定法** [7] によって得られる, **再ラベル進化系統樹** (relabeled phylogenetic tree) と**剪定進化系統樹** (trimmed phylogenetic tree) を使用する.

そして本論文では, 上記の 5 つのカーネルを用いて, 塩基配列と塩基配列位置を分類する. 塩基配列の分類では, A 型 H1N1 亜型インフルエンザウイルスに対してパンデミック前後と地域性を分類する. パンデミック前後の分類では, パンデミック後に発生した塩基配列

*連絡先：九州工業大学大学院情報工学府
〒 820-8502 福岡県飯塚市川津 680-4
E-mail: hamada@dumbo.ai.kyutech.ac.jp

を正例, 前に発生した塩基配列を負例として分類する. また, 地域性の分類では, 一つの地域の塩基配列を正例, 残りの地域の塩基配列を負例として分類する. 一方, 塩基配列位置の分類では, A 型 H3N2 亜型インフルエンザウイルスに対して, パッケージングシグナル位置に該当する位置を正例, そうでない位置を負例として分類する.

実験の結果, 塩基配列カーネルは, 塩基配列の分類には効果がないが, 塩基配列位置の分類には効果があり, 進化系統樹カーネルは, 塩基配列位置の分類には効果がないが, 塩基配列の分類には効果があることが分かった.

2 塩基配列カーネル

本節では塩基配列をベクトル, 重集合, 文字列とみなし, それぞれに対応するカーネルで, 塩基配列を分類するカーネルとして利用する. $U = A, C, G, T$ とする.

初めに, 塩基配列を U 上のベクトルとみなす. $A = (a_1, \ldots, a_n), B = (b_1, \ldots, b_n)$ (ただし, $a_i, b_i \in \Sigma$) を n 次元の配列 (ベクトル) とする. $\delta_1(a, b)$ を $a = b$ のとき 1, それ以外で 0 とし, $\delta_2(a, b)$ を $a = b$ のとき 1, (a, b) が (A, T),(T, A),(C, G),(G, C) のとき 1/2, それ以外で 0 とする. このとき, カーネル K_j を以下のように定義する.

$$K_j(A, B) = \frac{1}{n} \sum_{i=1}^{n} \delta_j(a_i, b_i).$$

次に, 塩基配列を U 上の重集合とみなす. 自然数 $\mathbf{N}$ に関して, $A \subseteq U \times \mathbf{N}$ を U の**重集合** (multiset) という. 重集合 A に対して, $(a, n) \in A$ の n を A における a の出現数といい, $\Gamma_A(a)$ と表す. なお, 重集合 $\{(a_i, n_i) \mid 1 \leq i \leq k\}$ を, a_i を n_i 個並べて $\{\underbrace{a_1, \ldots, a_1}_{n_1}, \ldots, \underbrace{a_k, \ldots, a_k}_{n_k}\}$ とも表す. また, $a \in U$ に対して $(a, 0) \in A$ となる要素は特に記述しない.

重集合 A, B に対して, $A \cap B = \{(a, n) \mid (a, n_A) \in A, (a, n_B) \in B, n = \min\{n_A, n_B\}\}$ と定義する. また, 重集合 A に対して, $|A| = \sum_{(a_i, n_i) \in A} n_i$ とする.

このとき, U の重集合上の**重集合直積カーネル** (multiset product kernel) $K_\times$ [10] と**重集合共通部分カーネル** (multiset intersection kernel) $K_\cap$ [10] を以下のように定義する. ここで, A,B は U 上の重集合である.

$$\begin{aligned} K_\times(A, B) &= \sum_{a \in U} \Gamma_A(a) \Gamma_B(a), \\ K_\cap(A, B) &= \sum_{a \in U} \min\{\Gamma_A(a), \Gamma_B(a)\}. \end{aligned}$$

最後に, 塩基配列を U 上の文字列としてみなす. $\Gamma_A(s)$ を A 中の部分文字列 s の出現した回数とする. 2 つの文字列に共通に出現する固定長の連続部分文字列の頻度を比較する**文字列カーネル** (spectrum string kernel) [5] を以下のように定義する.

$$K_{\mathrm{S}}^{\mathrm{n}}(A, B) = \sum_{\mathrm{s} \in \sum^{\mathrm{n}}} \Gamma_A(s) \Gamma_B(s).$$

3 進化系統樹カーネル

本節では, 進化系統樹, および, Hamada ら [3] が導入した**合致部分木マッピングカーネル** $K_{\mathrm{ASTM}}(T_1, T_2)$, および, **葉間パスカーネル** $K_{\mathrm{LP}}(T_1, T_2)$ を紹介する.

閉路を持たない連結無向グラフを**木** (tree) という. 木 $T = (V, E)$ に対して, $v \in V$ を $v \in T$, $|V|$ を $|T|$ と表す. また, 1 つのノード v を**根** (root) として選んだ木を**根つき木** (rooted tree) という.

根を r とする根つき木の任意のノードを v としたとき, v から r までの経路を $UP_r(v)$ と表す. $UP_r(v)$ 上で, v に隣接するノードを v の**親** (parent), $UP_r(v) - \{v\}$ を v の**先祖** (ancestors) という. また, u の親が v のとき, u は v の**子** (child) という. さらに, u と v の共通の先祖のうち, 最も u, v に近い先祖を**最近共通先祖** (least common ancestor) といい, $u \sqcup v$ と表す.

子を持たないノードを**葉** (leaf) といい, 葉でないノードを**内部ノード** (internal node) という. T の全ての葉の集合を $lv(T)$ と表し, 葉 v のラベルを $l(v)$ と表す. 葉のみにラベルがついている根付き木を**葉ラベル木** (leaf labeled tree) という. また, すべての内部ノードが 2 つの子を持つ木を**全二分木** (full binary tree) という. 本論文では, 無順序葉ラベル全二分木を**進化系統樹** (phylogenetic tree) という.

進化系統樹 T が与えられたとき, T のラベルの集合を $lv(T)$ とし, $L \subseteq lv(T)$ とする. このとき, T に以下の動作を適用する.

1. L のラベルを持つすべての葉から根への経路を求め, いずれの経路にも出現しない頂点を削除し, それらの頂点に接続している辺も削除する.

2. 子を一つしか持たない頂点を削除した後, その子を削除した頂点の親に接続する.

この二つの動作を適用して構成された木のことを, L によって制限された T の部分木といい, $T|L$ と表す. 二つの木 T_1 と T_2 に対して, $T_1|L_1$ と $T_2|L_2$ が同型となる木を T_1 と T_2 の**合致部分木** (agreement subtree) という.

次に, **マッピング** [11] によって合致部分木を特徴付けた**合致部分木マッピング**を導入する.

定義 1 (マッピング [11]) T_1 と T_2 を進化系統樹とする. T_1 と T_2 間のノードの一対一関係を $M \subseteq T_1 \times T_2$ とするとき, 以下の条件を満たした M を**マッピング**という.

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

合致部分木マッピングを定式化するために, T_1 と T_2 のマッピング M に対して, M^{lv} を $M \cap (lv(T_1) \times lv(T_2))$ と定義する.

定義 2 (合致部分木マッピング [3]) T_1 と T_2 を進化系統樹, M を T_1 と T_2 のマッピングとする. M が以下の条件を満たすとき, M を**合致部分木マッピング**という.

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. $\forall (v, w) \in M - M^{lv}$
$$\exists (v_1, w_1), (v_2, w_2) \in M^{lv}$$
$$\Big((v = v_1 \sqcup v_2) \wedge (w = w_1 \sqcup w_2)\Big).$$

定義 3 (合致部分木マッピングカーネル) 合致部分木マッピングカーネルは $K_{\mathrm{ASTM}}(T_1, T_2)$ といい, 木 T_1 と T_2 間のすべての合致部分木マッピングの数を表す.

定義 4 (葉間パスカーネル [6]) T_1と T_2 を進化系統樹とする. このとき, T_1と T_2 の**葉間パスカーネル** (leaf path kernel) $K_{LP}(T_1, T_2)$ を, 以下のように定義する.

$$K_{LP}(T_1, T_2) = \sum_{a \in L} \sum_{b \in L, b \geq a} \sum_{n=0}^{D} f_{T_1}(a, b, n) \cdot f_{T_2}(a, b, n).$$

ここで, $D = 2 \cdot \max\{dep(T_1), dep(T_2)\}$ である.

以降, 再ラベル進化系統樹を用いた合致部分木マッピングカーネルと葉間パスカーネルを, K^r_{ASTM}, K^r_{LP} と表し, 剪定進化系統樹を用いた合致部分木マッピングカーネルと葉間パスカーネルを, K^t_{ASTM}, K^t_{LP} と表す.

4 塩基配列の分類

本節では, 正例と負例に塩基配列のセットを分割する. その後, 進化系統樹カーネルでは, 正例の塩基配列と負例の塩基配列から, 二つの異なる進化系統樹を構成し, 使用する。よって, 正例から得られた再ラベル進化系統樹と剪定進化系統樹の数は, 負例から得られた再ラベル進化系統樹と剪定進化系統樹の数と等しくなり, その数は塩基配列長と一致する.

4.1 パンデミック前後の類別

パンデミック前後の類別では, NCBI データ [1] から得られた, A 型 H1N1 インフルエンザウイルスをパンデミック前の 2008 年の塩基配列 326 株を負例, パンデミック後の 2009 年の塩基配列 3344 株を正例とし, 塩基配列カーネルと進化系統樹カーネルを適用する. 塩基配列長は 895 である.

表 1 は, こうして得られたグラム行列に対して, LIBSVM を用いて 5 分割交差検証を行った結果である.

ここで, 塩基配列カーネルでは, 3670×3670 のグラム行列を使う. 一方, 進化系統樹カーネルでは, 両方に共通した位置でラベル付けされた進化系統樹だけを使用したので, 正例と負例ともに進化系統樹の数は 305 個となる. よって, 進化系統樹カーネルでは, 610×610 のグラム行列を使う. なお, 再ラベル進化系統樹に対しては, 合致部分木マッピングカーネルの値を計算できず, グラム行列を得ることができなかった.

表 1: パンデミック前後の分類結果.

	K_1, K_2	$K_\times$	$K_\cap$	$K^n_{\mathrm{S}} (n = 1, 2, 3, 4, 5)$	K^t_{ASTM}	K^t_{LP}	K^r_{LP}
F 値	1	0.999	0.999	1	0.911	0.915	1
AUC	1	0.999	0.999	1	0.951	0.866	1

表 1 より, 塩基配列カーネルと進化系統樹カーネルの両方で高精度で分類できていることが分かる.

次に, パンデミック前のデータとパンデミック後のデータからそれぞれ 200 株づつ抜き出すことで, データ数を揃えた上で, 塩基配列カーネルと進化系統樹カーネルを適用する.

ここで, 塩基配列カーネルでは, 200×200 のグラム行列を使う. 一方, 進化系統樹カーネルでは, 正例と負例の再ラベル進化系統樹の葉の数は 200 と同じになる.

表 2 は, こうして得られたグラム行列に対して, LIBSVM を用いて 5 分割交差検証を行った結果である.

表 2: データ数を揃えたパンデミック前後の分類結果.

	K_1, K_2	$K_\times$	$K_\cap$	$K^n_{\mathrm{S}} (n = 1, 2, 3, 4, 5)$	N^t_{ASTM}	K^r_{ASTM}	$K^t_{\mathrm{LP}}, K^r_{\mathrm{LP}}$
F 値	1	0.995	0.997	1	0.975	0.975	1
AUC	1	0.998	0.995	1	0.992	0.998	1

表 2 より, データ数を同じにしても, 塩基配列カーネルと進化系統樹カーネルの両方で高精度で分類できていることが分かる.

4.2 地域間の解析

地域間の解析には, NCBI データ [1] から得られた, A 型 H1N1 亜型インフルエンザウイルスの塩基配列データ 3670 株を地域別に分類して用いる. なお, 分類する地域は, アフリカ (AF), アジア (AS), ヨーロッパ (EU), 中東アジア (ME), 北アメリカ (NA), オセアニア (OC), 南アメリカ (SA) の 7 地域とする.

表 3 は, 地域別の塩基配列データの株数 (#NS) と, 進化系統樹 (再ラベル進化系統樹, および, 剪定進化系統樹) の本数 (#PS) と, それぞれの全体に占める割合である.

表 3: 地域別の塩基配列データの株数と進化系統樹の本数と割合.

	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	
#PS	289	593	487	311	538	290	344	2852
%	10.13	20.79	17.08	10.90	18.86	10.17	12.06	

表 3 より, 塩基配列カーネルでは 3670×3670 のグラム行列となり, 例えばアフリカを分類する場合は, アフリカの 61 株を正例, それ以外の 3609 株を負例とする. このように各地域を正例とした 7 地域のグラム行列を得る.

また, 表 3 より, 進化系統樹カーネルでは 2852×2852 のグラム行列となり, 例えばアフリカを分類する場合は, アフリカの 289 本を正例, それ以外の 2563 本を負例とする. このように各地域を正例とした 7 地域のグラム行列を得る. ここで, 再ラベル進化系統樹に対しては, 合致部分木マッピングカーネルの値を計算できず, グラム行列を得ることができなかった.

表 4 はこのようにして得られた, 塩基配列カーネルと進化系統樹カーネルの各地域のグラム行列に対して, LIBSVM を用いて 5 分割交差検証を行い, 分類精度を比較した結果である.

表 4 より, 進化系統樹カーネルでの分類には成功しているが, 塩基配列カーネルでは分類に失敗していることが分かる.

次に, 表 4 より, 両方とも高い分類結果を出した, 進化系統樹カーネルを用いて, 二つの地域の組合わせで片方を正例, もう片方を負例とて分類する. 例えば, アフリカとアジアの分類を行う場合, アフリカの 289 本を正例とし, アジアの 593 本 を負例とすることで, 882×882 のグラム行列となる. ここで, 再ラベル進化系統樹に対

表 4: 地域の分類結果.

		AF	AS	EU	ME	NA	OC	SA
K_1	F 値	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 値	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 値	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 値	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 値	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 値	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 値	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 値	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 値	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_{ASTM}^t	F 値	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 値	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 値	1	1	1	0.998	1	1	1
	AUC	1	1	1	0.999	1	1	1

しては, 合致部分木マッピングカーネルの値を計算できず, グラム行列を得ることができなかった.

表 5 から表 7 は, 進化系統樹カーネルを適用して得られたグラム行列を, LIBSVM を用いて 5 分割交差検証を行った結果である.

表 5: 合致部分木マッピングカーネルを用いた地域の分類結果.

K_{ASTM}^t		AS	EU	ME	NA	OC	SA
AF	F 値	0.967	1	0.940	0.989	0.949	0.994
	AUC	0.991	1	0.940	0.989	0.949	0.994
AS	F 値		0.963	0.982	0.914	0.987	0.885
	AUC		0.990	0.991	0.945	0.994	0.871
EU	F 値			0.998	0.944	0.998	0.989
	AUC			1	0.971	1	0.999
ME	F 値				0.980	0.756	0.998
	AUC				0.997	0.771	0.999
NA	F 値					0.996	0.937
	AUC					0.999	0.969
OC	F 値						1
	AUC						1

表 5 から表 7 より, いずれの進化系統樹カーネルでも, かなり高精度で分類できていることが分かる. 特に, 再ラベル進化系統樹を用いた葉間パスカーネルでは完全に分類できていることが分かる.

5 塩基配列位置の分類

本節では, 一組の塩基配列をパッケージングシグナル位置に該当する位置, 正位置とそうでない位置, 負位置に分割する. その後, 進化系統樹カーネルでは, 塩基配列全体から再構成された一本の進化系統樹を使用する.

表 6: 葉間パスカーネルを用いた地域の分類結果 (剪定進化系統樹).

K^t_{LP}		AS	EU	ME	NA	OC	SA
AF	F 値	0.944	1	0.914	0.975	0.956	0.993
	AUC	0.985	1	0.925	0.995	0.967	0.999
AS	F 値		0.996	0.975	0.865	0.984	0.910
	AUC		0.999	0.984	0.921	0.994	0.936
EU	F 値			1	0.960	1	0.993
	AUC			1	0.988	1	0.999
ME	F 値				0.977	0.920	0.998
	AUC				0.991	0.934	0.999
NA	F 値					0.992	0.932
	AUC					0.997	0.954
OC	F 値						0.998
	AUC						1

表 7: 葉間パスカーネルを用いた地域の分類結果 (再ラベル進化系統樹).

K^r_{LP}		AS	EU	ME	NA	OC	SA
AF	F 値, AUC	1	1	1	1	1	1
AS	F 値, AUC		1	1	1	1	1
EU	F 値, AUC			1	1	1	1
ME	F 値, AUC				1	1	1
NA	F 値, AUC					1	1
OC	F 値, AUC						1

よって, 正位置でラベル付けされた進化系統樹と負位置でラベル付けされた進化系統樹の木の構造は, 葉のラベルが違うだけで一致する.

5.1 パッケージングシグナル

A 型インフルエンザウイルスは, 粒子内部に PB2, PB1, PA, HA, NP, NA, MP(もしくは M), NS という 8 種類に**分節化された** RNA を持つ. インフルエンザウイルスの粒子形成過程では, ウイルスの遺伝子情報は, 8 種類の RNA 分節により感染細胞から次の感染細胞へと運ばれる [4]. インフルエンザウイルスの粒子形成過程において, ウイルス 1 個から完全な子孫ウイルスが複製されるためには, 各ウイルス粒子に 8 種類すべての RNA 分節が含まれなければならない.

インフルエンザウイルスの粒子形成では, 8 種類 8 本の RNA 分節が宿主細胞内で選択的に集合し, すべての分節が 1 本ずつ子孫ウイルスの粒子内に取り込まれる傾向がある. この粒子形成で, 集合の背後に何らかの規則があると仮定するモデルを specific model (もしくは selective-incorporation model) という. この specific model の粒子形成で RNA 分節の集合に必要な塩基配列を**パッケージングシグナル** [4] という.

表 8 は, リバースジェネティックスによって確認されている各 RNA 分節のパッケージングシグナルの位置である. リバースジェネティックスとは, ウイルスを人工的に合成する技術のことである. ここで, (5' xx-)yy と (3' xx-)yy の xx は, 5' および 3' の UTR(非解釈領域) の塩基数であり, yy はそれに続く塩基数を表す. また, (5')xx–yy は, 5' を始点とした xx 番目から yy 番目の塩基を表す. さらに, NCBI の欄は, 本論文で用いる NCBI データ [1] における番号である. また, パッケージングシグナル位置を正位置, それ以外の位置を負位置とし, 位置の欄は, 分節ごとの正位置の総数 (+), 負位置 (−) の総数である.

表 8: RNA 分節のパッケージングシグナルの位置 [4]. ε は NCBI に該当番号が存在しないことを表す.

分節	位置	NCBI	正位置 (+)	負位置 (−)
PB2 (2341)	(5' 34-)80 (5')2209–2304	35–114 2209–2304	174	2167
PB1 (2341)	(5' 43-)66, (3' 24-)40 (5' 43-)120, (3' 24-)12 (5' 43-)12, (3' 24-)120 (5')2256–2279	38–66, 2277–2299 38–163, ε 38–55, 2197–2299 2256–2279	227	2114
PA (2233)	(5' 58-)66, (3' 24-)40 (5')691–731 (5')742–767 (5')2094–2156	38–124, 2169–2176 691–731 742–767 2094–2156	220	2013
HA (1778)	(5' 45-)80, (3' 33-)9 (5')1659–1671	38–125, ε 1659–1671	99	1679
NP (1565)	(5' 45-)120, (3' 23-)60	46–165, 1482–1526	163	1402
NA (1413)	(5' 28-)157, (3' 19-)183 (3' 19-)21 (critical)	35–185, 1211–1413 1373–1413	352	1061
MP (1027)	(5' 6-)29	ε	-	-
NS (890)	(3' 12-)30 (5')16–56 (5')841–870	ε 36–56 ε	20	870

5.2 パッケージングシグナル位置解析

本節では, A 型 H3N2 インフルエンザウイルスの 7 つの RNA 分節の塩基配列それぞれから, パッケージングシグナル位置を正例とし, そうでない位置を負例とし, 塩基配列カーネルに適用する. 例えば, PB1 分節であれば, 表 8 より, 正例は 227, 負例は 2114 のとなり, 2341×2341 のグラム行列となる. 進化系統樹カーネルでは, 縦がすべて同じ塩基である位置を削除した後, パッケージングシグナル位置でラベル付けされた進化系統樹を正例, それ以外を負例とし, 適用する.

表 9 は, そうして得られたグラム行列を, LIBSVM を用いて 5 分割交差検証を行い, 分類精度を比較した結果である.

表 9: パッケージングシグナル位置の分類結果.

		PB2	PB1	PA	HA	NP	NA	NS
塩基配列カーネル	F 値	1	1	1	1	1	1	1
	AUC	1	1	1	1	1	1	1
K^t_{ASTM}	F 値	0	0	0	0	0	0.045	0
	AUC	0.462	0.494	0.515	0.527	0.474	0.486	0.493
K^t_{LP}	F 値	0.092	0.080	0.154	0.058	0.131	0.278	0
	AUC	0.530	0.431	0.513	0.479	0.529	0.479	0.481

表 9 より, 塩基配列カーネルでの分類は完全に成功しているが, 進化系統樹カーネルでの分類に失敗していることが分かる.

次に表 9 より, 完全な分類結果を出した塩基配列カーネルを用いて, 縦の位置が同塩基である箇所を削除し, 正例と負例でデータ数を等しくしたデータを使用してグラム行列を作成した. 各分節の配列長はそれぞれ, PB2 が 174 から 150, PB1 が 227 から 113, PA が 220 から 87, HA が 99 から 77, NP が 163 から 64, NA が 352 から 205, NS が 20 から 11 となった.

表 10 は, そうして得られたグラム行列を, LIBSVM を用いて 5 分割交差検証を行い, 分類精度を比較した結果である. ここで, 位置を選ぶことは, 部分文字列が意味をなさないため, 文字列カーネルは適用しない.

表 10: 配列長を揃えたパッケージングシグナル位置の分類結果.

		PB2	PB1	PA	HA	NP	NA	NS
K_1	F 値	0.999	1	1	1	1	0.999	1
	AUC	1	1	1	1	1	1	1
K_2	F 値	0.999	1	1	1	1	0.999	1
	AUC	1	1	1	1	1	1	1
$K_\times$	F 値	1	1	0.999	1	0.998	1	0.994
	AUC	1	1	1	1	0.999	1	0.995
$K_\cap$	F 値	1	1	0.999	1	0.997	1	0.995
	AUC	1	1	1	1	0.999	1	0.996

表 10 より, すべての塩基配列カーネルにおいて高精度で分類できていることが分かる.

6 まとめ

本論文では, 塩基配列をそのまま用いる塩基配列カーネルと, 塩基配列から得られる進化系統樹を用いる進化系統樹カーネルによってインフルエンザウイルスの塩基配列と塩基配列位置を分類した.

塩基配列の分類のうち, パンデミック前後を分類する実験はすべてのカーネルで高精度だった. また, 地域間を分類する実験では, 塩基配列カーネルは分類ができていなかったが, 進化系統樹カーネルでは高精度で分類できたところが多く, 特に再ラベル進化系統樹では, すべての地域で高精度で分類することができた.

一方, 塩基配列位置の分類では, 進化系統樹カーネルは分類ができなかったが, 塩基配列カーネルでは高精度で分類することができた.

この理由として, 塩基配列の分類では正例と負例それぞれで再構成される進化系統樹が分類の背景知識として分類に有効だったことに対して, 塩基配列位置の分類では, 進化系統樹は 1 つしか再構成されず, それが分類を妨げたことが考えられる.

今後の課題として, H1N1 のパッケージングシグナル位置の分類や, H3N2 の地域間の分類を行うこと. また, アミノ酸配列等, 別の塩基配列へカーネルを用いて分類を行うことが挙げられる.

参考文献

[1] T. Bao, P. Bolotov, D. Dernovoy, B. Kiryutin, L. Zaslavsky, T. Tatusova, J. Ostell, D. Lipman: *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/`.

[2] C.-C. Chang, C.-J. Lin: *LIBSVM – A library for support vector machine* (version 3.17), 2013 `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

[3] I. Hamada, T. Shimada, K. Hirata, T. Kuboyama: *agreement subtree mapping kernel for phylogenetic trees*, Proc. DDS 2013, 1-8, 2013.

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

[5] C.S. Leslie, E. Eskin, W.S. Noble: *The spectrum kernel: A string kernel for SVM protein classification*, Proc. PSB 2002, 566-575, 2002.

[6] 島田　昂治, 濵田 一青, 平田耕一: 進化系統樹の葉間パスカーネルによるパンデミック解析, SIG-FPAI-B302, 7-12, 2013

[7] S. Makino, T. Shimada, K. Hirata, K. Yonezawa, K. Ito: *A trim distance between positions in nucleotide sequences*, Proc. DS 2012, LNAI **2569**, 81–94, 2012.

[8] K. Shin, T. Kuboyama: *Kernels based on distributions of agreement subtrees*, Proc. AI 2008, LNAI **5360**, 236–246, 2008.

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

[10] T. Gärtner: *Kernels for structured data*, World Scientific, 2008.

[11] K.-C. Tai, *The tree-to-tree correction problem*, J. ACM **26**, 422–433, 1979.