人工知能学会研究会資料
SIG-FPAI-B501-08

A 型インフルエンザウイルスの RNA 分節における極大対合ヘアピンループ探索

Finding Maximal Kissing Hairpins through RNA Segments in Influenza A Viruses

嶋村 翔[1]　平田 耕一[2]
Sho Shimamura[1]　Kouichi Hirata[2]

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

Abstract: In this paper, first we formulate a *kissing hairpin* between two RNA sequences as RNA tertiary structures kissing two hairpin loops for every RNA sequence. Then, we design the method of finding all of the *maximal* kissing hairpins between two RNA sequences, where a kissing hairpin is maximal if the length of its kissing part is maximal. Finally, by connecting maximal kissing hairpins with the same kissing part between two RNA segments, we give experimental results of finding maximal kissing hairpins through 8 RNA segments in influenza A (H3N2) viruses.

1　はじめに

A 型インフルエンザウイルスは, PB2, PB1, PA, HA, NP, NA, NS という 8 種類の RNA 分節から構成されており, これはウイルス粒子化における選択的な**パッケージング** (*packaging*) により, 8 分節から 1 本ずつ集まると考えられている. それら 8 本を選択的にパッケージングする背景となる**塩基**を**パッケージングシグナル** (*packaging signal*) という.

このパッケージングシグナルの分析にはこれまで, 文字列アラインメント [4, 10], 共変異 [11], カーネル [7] などの手法が用いられてきたが, これらの方法はパッケージングシグナルを特徴付けるためにはまだ不十分であった. 一方, Gavazzi ら [5] は, RNA 分節のうち PB1 と NS が図 1 のような対合ヘアピンループを形成していることを指摘している.

ここで, **ヘアピンループ** (*hairpin loops*) [1, 3, 12] とは, 塩基対を畳み込んだ RNA 二次構造の一つであり, **対合ヘアピンループ** (*kissing hairpin loops*) [1] とは 2 つのヘアピンループの一部が反転相補となって塩基対により結合している RNA 三次結合の一つである.

そこで, 本論文では A 型インフルエンザイウイルスの 8 つの RNA 分節間の対合ヘアピンループを抽出する.

まず最初に, *Nussinov* **アルゴリズム** [3, 12] を簡略化し, それにより, RNA 配列からヘアピンループをすべて求める. 次に, 2 つの RNA 配列から得られたヘアピンループを比較することにより, それらの**極大**対合ヘアピンループをすべて抽出する. ここで, **極大**対合ヘアピンループとは対合部の長さがそれ以上の長さを持つ対合部がないことをいう. したがって本論文は, 図 1 のステム部にある A のようなミスマッチとなる塩基は考えない.

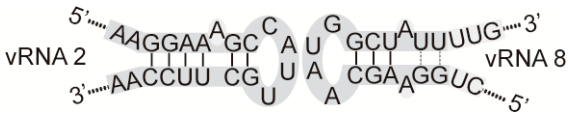

図 1: 対合ヘアピンループ [5]. ここで, vRNA2, vRNA8 はそれぞれ 5' から始まり 3' への RNA 分節 PB1 と NS を意味している.

最後に, NCBI [2] から提供されている A 型 (H3N2) インフルエンザウイルスが持つ 8 本の RNA 分節を対象に, 隣接する分節間の極大対合ヘアピンループを探索する. そして, それらの極大対合ヘアピンループを接続することにより, 8 本の RNA 分節すべてに渡る極大対合ヘアピンループと, パッケージングシグナル位置を比較する.

2 ヘアピンループと対合ヘアピンループ

Σ を アルファベット $\{\mathrm{A},\mathrm{U},\mathrm{G},\mathrm{C}\}$ とする．$a \in \Sigma$ を**塩基**, $s \in \Sigma^*$ を RNA **配列**という．また, $\{\mathrm{A},\mathrm{U}\}$, $\{\mathrm{G},\mathrm{C}\}$ の組み合わせを**塩基対**といい, A (U, G, C) に対する U (A, C, G) を**相補**という．$a \in \Sigma$ に対し, a の相補を a^c と表す.

RNA 配列 s に対して、s の長さを $|s|$ と表す．また, $1 \leq i \leq |s|$ の i に対して, s の i 番目の記号を $s[i]$ と表し, i から j までの s の部分文字列 $s[i]\cdots s[j]$ を $s[i;j]$(ただし $1 \leq i \leq j \leq |s|$), $|s| = n$ であるような文字列を $s[1;n]$ で表す.

定義 1 (RNA 二次構造, ヘアピンループ) RNA 配列 s に対して, $P \subseteq \{1,\ldots,|s|\} \times \{1,\ldots,|s|\}$ とする．P が以下 3 つの条件を満たすとき, P を **RNA 二次構造** (*RNA secondary structure*) という.

1. $\forall (i,j) \in P \left(s[i] = s[j]^c\right)$.
2. $\forall (i,j) \in P \left(j > i+1\right)$.
3. $\forall (i,j),(i',j') \in P \left(i = i' \leftrightarrow j = j'\right)$.

また, RNA 二次構造 P がさらに以下の条件を満たすとき, P を**ヘアピンループ**という.

1. $\forall (i,j),(i',j') \in P$
   $\Big((i+1 < i') \wedge (j' < j-1) \rightarrow$
   $\exists (i'',j'') \in A\Big((i < i'' < j') \wedge (j' < j'' < j)\Big)\Big)$.
2. $\forall (i,j) \in P \left((i+1,j') \in A \rightarrow j' = j-1\right)$.
3. $\forall (i,j) \in P \left((i',j-1) \in A \rightarrow i' = i+1\right)$.

注意 1 $\#P = d$(ここで $\#P$ は集合 P の濃度を表す) となるヘアピンループ P に対して, 次の条件を満たす i,$j(i+1<j)$ が常に存在する.

$$P = \left\{ \begin{array}{l} (i,j),(i+1,j-1),\ldots, \\ (i+d-1,j-d+1) \end{array} \middle| \, i+d < j-d+1 \right\}$$
$$= \left\{ (i+k,j-k) \middle| \begin{array}{l} 0 \leq k \leq d-1, \\ i+d < j-d+1 \end{array} \right\}.$$

そこで, このようなヘアピンループ P を $\langle i,j;d\rangle$ と表す．このとき, $s[i;i+d-1]$ と $s[j-d+1;j]$ をステム部, $s[i+d;j-d]$ を**ループ部**といい, $|s[i;i+d-1]| = |s[j-d+1;j]| = d$ を**ステム長** (*stem length*), $|s[i+d;j-d]| = j-i-2d+1$ を**ループ長** (*loop length*), $j-i$ を**全長** (*total length*) という.

2 つのヘアピンループ $P_1 = \langle i,j;d\rangle$ と $P_2 = \langle i',j';d'\rangle$ に対して, $i'+k = i$, $j'-k = j$, $d'-k = d$ となるような $k > 0$ が存在するとき, P_1 は P_2 を**含む**, もしくは もしくは P_2 は P_2 によって**含まれる**という．また, ヘアピンループ P を含むヘアピンループが存在しないとき, P は**極大**であるという.

例 1 s_1 = AGAACAUUCUCACCGU に対して, ステム長が 1 よりも大きい s のヘアピンループは $P_1 = \langle 1,10;4\rangle$, $P_2 = \langle 2,9;3\rangle$, $P_3 = \langle 3,8;2\rangle$, $P_4 = \langle 2,11;2\rangle$, $P_5 = \langle 4,16;2\rangle$, $P_6 = \langle 12,16;2\rangle$ である．図 2 は P_1, P_2, P_3, P_6 を上側の弧で, P_4, P_5 を下側の弧で表している．また, P_1, P_4, P_5, P_6 は極大であり, P_2, P_3 は極大ではない.

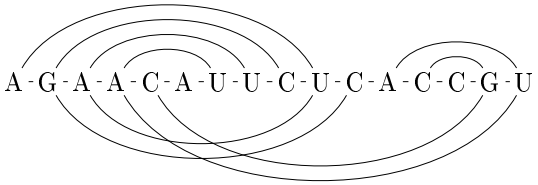

図 2: 例 1 の s_1 に対する長さが 1 以上のヘアピンループ

$s = s[1]\cdots s[n] \in \Sigma^*$ を RNA 配列とする．s の**反転** (*reverse*) $s[n]\cdots s[1]$ を s^{-1} と表す．また, s の**相補** (*complement*)$s[1]^c\cdots s[n]^c$ を s^c と表す．このとき, $(s^c)^{-1} = (s^{-1})^c$ は明らかである．長さが n の 2 つの配列 s_1, s_2 に対して, $s_1 = (s_2^{-1})^c$, つまり, $s_1[1]\cdots s_1[n] = s_2[n]^c\cdots s_2[1]^c$ が成り立つとき, s_2 は s_1 の**反転相補** (*reverse complement*) といい, $s_1 = (s_2)^{-c}$ または $s_2 = (s_1)^{-c}$ と示す. 例えば, AUGGCU の反転相補は $\mathrm{AUGGCU}^{-c} = \mathrm{UCGGUA}^c = \mathrm{AGCCAU}$ である.

定義 2 (対合ヘアピンループ) s_1 と s_2 を RNA 配列とし, P_1 を s_1 のヘアピンループ, P_2 を s_2 のヘアピンループとする．$s_1' = (s_2')^{-c}$ となる P_1 のループ部の部分列 s_1' と, P_2 のループ部の部分列 s_2' が存在するとき, s_1 と s_2 は P_1 と P_2 で, **対合ヘアピンループ**を形成しているという．s_1', s_2' の**対合部**と長さを**対合長** (*kissing length*) とする．このとき, P_1, P_2 を $\boldsymbol{s_1}$, $\boldsymbol{s_2}$ **間の対合ヘアピンループ**とし, さらに, 対合長が極大であり, どの対合部にも含まれない対合ヘアピンループを**極大**であるといい, 極大となる対合ヘアピンループを極大対合ヘアピンループという.

例 2 例 1 の RNA 配列 $s_1 =$ AGAACAUUCUCACCGU と RNA 配列 $s_2 =$ UAGGAGAAUGUACCGA を考える．このとき，s_1 には $\langle 2, 11; 2\rangle$，s_2 には $\langle 1, 12; 2\rangle$ となるヘアピンループが存在する．

ここで，$\langle 2, 11; 2\rangle$ のループ部は $s_1[5; 9] =$ CAUUC となり，$\langle 1, 12; 3\rangle$ のループ部は $s_2[6; 10] =$ GAAUG となる． $s_2[6; 10]^{-c} =$ GAAUG$^{-c} =$ GUAAG$^{c} =$ CAUUC $= s_1[5; 9]$ となるため，$s_2[6; 10]$ は $s_1[5; 9]$ の反転相補である．

したがって，s_1 と s_2 は $s_1[5; 9]$, $s_2[6; 10]$ で対合部を形成する s_1, s_2 間の対合ヘアピンループである．この対合ヘアピンループは $\langle 2, 11; 2\rangle$,$\langle 5, 9\rangle$,$\langle 1, 12; 2\rangle$,$\langle 10, 6\rangle$ で表すことができ，同様に $\langle 2, 11; 2\rangle$,$\langle 9, 5\rangle$,$\langle 1, 12; 2\rangle$,$\langle 6, 10\rangle$ でも同等に表せる．

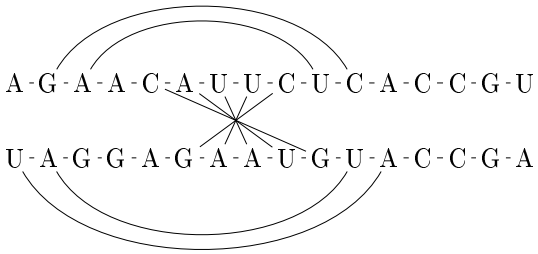

図 3: 例 2 s_1, s_2 の対合ヘアピンループ．

定義 1, 2 より，本論文のヘアピンループと対合ヘアピンループは**ミスマッチ** (*mismatch*) を含まないものとする．

3 極大対合ヘアピンループ探索アルゴリズム

Σ における塩基 x, y に対して，次のように $\delta(x, y)$ を定義する．

$$\delta(x, y) = \begin{cases} 1, & \text{if } \{x, y\} = \{\mathrm{A}, \mathrm{U}\}, \{\mathrm{G}, \mathrm{C}\}, \\ 0, & \text{それ以外.} \end{cases}$$

このとき Nussinov アルゴリズム [3, 12] を簡略化することにより，構築するすべての i, j $(1 \le i \le j \le n)$ に対して，以下のテーブル $W[i, j]$ を構築し，長さ n の RNA 配列 s に含まれるすべてのヘアピンループが得られる．

$$\begin{cases} W[i, i] = W[i, i+1] = 0, \\ W[i, j] = \delta(s[i], s[j]) \cdot \Big(W[i+1, j-1] + \delta(s[i], s[j])\Big). \end{cases}$$

$W[i, j]$ テーブルを利用し，RNA 配列 s から $\langle i, j; d\rangle$ で表されるすべてのヘアピンループを抽出するアルゴリズムを設計する．

```
procedure HAIRPIN(s)
    /* s = s[1 ··· n] */
1   H ← ∅;
2   for i = 1 to n do
3       W[i, i] ← 0; W[i, i + 1] ← 0;
4   for d = 2 to n − 1 do
5       for i = 1 to n − d do
6           j ← i + d; W[i, j] ← δ(s[i], s[j]) ·
            (W[i + 1, j − 1] + δ(s[i], s[j]));
7           if W[i, j] ≠ 0 then
8               H ← H ∪ {⟨i, j; W[i, j]⟩};
9   return H;
```

アルゴリズム 1: HAIRPIN.

定理 1 $s \in \Sigma^*$ を $|s| = n$ となる RNA 配列とする．このとき，アルゴリズム HAIRPIN(s) は s に含まれるすべてのヘアピンループを $O(n^2)$ 時間 $O(k)$ 領域で正しく抽出することができる．ここで k は s に含まれる極大となるヘアピンループの数とする．

[証明] $s[i], s[j]$ が塩基対でないとき，(i, j) はヘアピンループに含まれず，$W[i, j]$ の値は 0 となる．それ以外は塩基対であるとする．$s[i+1]$ と $s[j-1]$ が塩基対ならば，$(i+1, j-1)$ を含むヘアピンループ P であり，P から $P \cup \{(i, j)\}$ である．これは $W[i, j] = W[i+1, j-1]+1$ であることを意味する．したがって，δ を利用し，$W[i, j]$ の値を式 $\delta(s[i], s[j]) \cdot \Big(W[i+1, j-1] + \delta(s[i], s[j])\Big)$ として計算することができる．

アルゴリズム HAIRPIN はヘアピンループをすべて抽出することが目的であり，Nussinov アルゴリズムの目的のひとつであるヘアピンループの極大化を考慮しない．そのため，アルゴリズム HAIPRPIN の 5 行目は Nussinov アルゴリズム [3, 12] と比較すると容易である．

アルゴリズム HAIRPIN は $O(n^2)$ 領域からなる $W[i, j]$ テーブル $(1 \le i \le j \le n)$ を参照するため，$O(n^2)$ 時間で実行できることが明らかである．さらに，極大となるヘアピンループ $\langle i, j; d\rangle$ から，それに含まれるすべてのヘアピンループ $\langle i+k, j-k; d-k\rangle (1 \le k \le d-1)$ を抽出することができる．したがって，$\langle i, j; 1\rangle$ を更新することにより極大のヘアピンループを抽出することができる．そのため，アルゴリズム HAIRPIN の計算領域は，$O(k)$ で十分である． □

H_1 と H_2 は，それぞれアルゴリズム HAIRPIN を用いて s_1, s_2 からとり出したすべてのヘアピンループの集

合とする．アルゴリズム KISSING は，H_1 のすべてのヘアピンループ P_1 と H_2 のすべてのヘアピンループ P_2 から極大の対合ヘアピンループをすべて探索するアルゴリズムである．アルゴリズム KISSING の出力は対合ヘアピンループの集合 KH である．

```
procedure KISSING(s_1, s_2)
    /* P_1 = ⟨i_1, j_1; d_1⟩, P_2 = ⟨i_2, j_2; d_2⟩ */
1   H_1 ←HAIRPIN(s_1); H_2 ←HAIRPIN(s_2);
    m ← 1;
2   foreach (P_1, P_2) ∈ H_1 × H_2 do
3       l_1 ← j_1 − i_1 − 2d_1 + 1;
        l_2 ← j_2 − i_2 − 2d_2 + 1;
4       for e_1 = 0 to l_1 do
5           i'_1 ← i_1 + d_1 + e_1 + 1;
            j'_1 ← j_1 − d_1 − e_1 − 1;
6           for e_2 = 0 to l_2 do
7               i'_2 ← i_2 + d_2 + e_2 + 1;
                j'_2 ← j_2 − d_2 − e_2 − 1;
8               if δ(s_1[i'_1], s_2[j'_2]) = 1 then
9                   k ← 1; i''_1 ← i'_1 + k;
                    j''_2 ← j'_2 − k;
10                  while δ(s_1[i''_1], s_2[j''_2]) = 1 do
11                      k++; i''_1++; j''_2−−;
12                  KH[m] ←
                    (P_1, ⟨i'_1, i''_1 − 1⟩, P_2, ⟨j'_2, j''_2 + 1⟩);
13              else if δ(s_1[j'_1], s_2[i'_2]) = 1 then
14                  k ← 1; j''_1 ← j'_1 − k;
                    i''_2 ← i'_2 + k;
15                  while δ(s_1[j''_1], s_2[i''_2]) = 1 do
16                      k++; j''_1−−; i''_2++;
17                  KH[m] ←
                    (P_1, ⟨j'_1, j''_1 + 1⟩, P_2, ⟨i'_2, i''_2 − 1⟩);
18              m++;
19      return KH;
```

アルゴリズム 2: KISSING.

定理 2 $|s_1| = n$, $|s_2| = m$ とする．このとき，アルゴリズム KISSING(s_1, s_2) はすべての s_1, s_2 間の極大対合ヘアピンループを $O(n^2m^2)$ 時間で抽出することができる．

本論文では，8 分節間の組み合わせの数が膨大となるため，隣接するペア (PB2 と PB1, PB1 と PA, PA と HA, HA と NP, NP と NA, NA と MP, MP と NS) を対象とし，対合ヘアピンループを抽出するアルゴリズム COMMON を設計する．また，8 本の RNA 分節に渡って極大対合ヘアピンループを構成する対合ヘアピンループの集合を *8* **分節間対合ヘアピンループ**という．

ここで，$\#S_j = L$ である RNA 分節 S の j 番目 $(1 \leq j \leq 8)$ を S_j で表す．i 番目の RNA 配列の j 番目の RNA 分節を $S_j[i]$ $(1 \leq i \leq L)$ で表す．

また，i 番目の RNA 配列に対してアルゴリズム COMMON の 5 行目から 8 行目はアルゴリズム COMMON の 7 行目で，全ての j $(1 \leq j \leq 6)$ に対して $P_2^j = P_1^{j+1}$ であるかを確認することにより，共通の対合ヘアピンループ $kh_j = (P_1^j, I_1^j, P_2^j, I_2^j)$ を探索している．そして，アルゴリズム COMMON では，RNA 配列 i 番目から見つかった 8 分節間対合ヘアピンループを $CKH[i]$ として格納している．

```
procedure
COMMON(S_1, S_2, S_3, S_4, S_5, S_6, S_7, S_8)
    /* S_1: PB2, S_2: PB1, S_3: PA, S_4: HA, S_5
    :NP, S_6: NA, S_7: MP, S_8: NS */
    /* L : the number of sequences for every
    segment (L = 1560) */
1   for i = 1 to L do
2       CKH[i] ← ∅;
3       for j = 1 to 7 do
4           KH_j[i] ← KISSING(S_j[i], S_{j+1}[i]);
5       foreach
        (kh_1, kh_2, kh_3, kh_4, kh_5, kh_6, kh_7) ∈
        KH_1[i] × KH_2[i] × KH_3[i] × KH_4[i] ×
        KH_5[i] × KH_6[i] × KH_7[i] do
6           /* kh_j = (P_1^j, I_1^j, P_2^j, I_2^j) (1 ≤ j ≤ 7) */
7           if (P_2^1, P_2^2, P_2^3, P_2^4, P_2^5, P_2^6) =
            (P_1^2, P_1^3, P_1^4, P_1^5, P_1^6, P_1^7) then
8               CKH[i] ← CKH[i] ∪
                {(kh_1, kh_2, kh_3, kh_4, kh_5, kh_6, kh_7)};
9   C ← ∅;
10  foreach ckh ∈ CKH[1] do
11      freq ← 1;
12      for i = 2 to L do
13          if ckh ∈ CKH[i] then
14              freq++;
15  C ← C ∪ {(ckh, freq)};
16  return C;
```

アルゴリズム 3: COMMON.

定理 1, 2 より 2 つの配列間の対合ヘアピンループを効率的にすべて抽出することができる．一方，アルゴリズム COMMON の計算時間は見つかった対合ヘアピンループの数に依存する．配列の組み合わせ数が増えると計算時間が非常に大きくなる．

このような困難を回避するために, ヘアピンループ $\langle i, j; d\rangle$ に対して閾値 D, L, H を, 対合ヘアピンループ $(\langle i_1, j_1; d_1\rangle, \langle I_1, I_1+k\rangle, \langle i_2, j_2; d_2\rangle, \langle J_2, J_2-k\rangle)$ に対して閾値 K を導入する.

D : 最小ステム長 (最小値: d),
L : 最小ループ長 (最小値: $j-i-2d+1$),
H : 最大全長 (最大値: $j-1$),
K : 最小対合長 (最小値: k).

閾値 D, L, K により, それぞれステム長, ループ長, 対合長の小さい対合ヘアピンループを処理しない. 一方で, 閾値 H を導入することで対合ヘアピンループの局所性を保障する. これらの閾値はアルゴリズム Hairpin 及び Kissing に容易に導入することができる.

4 実験結果

本節では, NCBI [2] から提供されている A 型 (H3N2) インフルエンザウイルスの 8 本の RNA 分節に対して第 3 節のアルゴリズムを用いて 8 分節間対合ヘアピンループを抽出する.

NCBI から提供されている RNA 配列の数は 1560 株あり, 各分節の長さはそれぞれ PB2 は 2274, PB1 は 2262, PA は 2145, HA は 1693, NP は 1481, NA は 1382, MP は 965, NS は 798 である.

RNA 配列 1560 株の隣接した分節間の対合ヘアピンループを抽出し, 抽出した対合ヘアピンループのヘアピンループが同じものを接続して 8 分節間対合ヘアピンループを構成する. このとき, 1560 株に含まれる 8 分節間対合ヘアピンループの数を**頻度**と呼ぶ.

閾値 D, L, K を 4 に設定し, H を 10 ずつ増加させ, 8 分節間対合ヘアピンループの抽出すると, H が 20 を超えたとき, 8 分節間対合ヘアピンループが出力された.

そこで, 閾値 $(4,4,20,4)$ を起点とし, 閾値 D, L, K を 1 ずつ, H を 30 ずつ増加させて 8 分節間対合ヘアピンループの抽出を行った. 閾値 D, L, K の値が 7 以上となったとき, 8 分節間対合ヘアピンループが抽出できなくなり, 閾値 $(4,4,20,4)$, $(5,5,50,5)$, $(6,6,80,6)$ では頻度が 50%を超えていた. そこで, 閾値 $(5,5,50,5)$ に着目する.

表 1 は閾値 $(5,5,50,5)$ において最大頻度で見つかった分節間対合ヘアピンループを示す.

分節間対合ヘアピンループが最初に見つかった時の頻度は, 1353 個であった. 表 1 でいう PB2-PB1 と NA-MP に頻度 1353 で対合ヘアピンループが発生し, それにより 8 分節間対合ヘアピンループが構成された.

本節の最後で, RNA 分節におけるパッケージングシグナルの位置を含む対合ヘアピンループに焦点を当てる.

表 1: 閾値 $(5,5,50,5)$ において最大頻度で見つかった分節間対合ヘアピンループ

RNA	hairpin loop	kissing part	hairpin loop	kissing part	freq.
PB2–PB1	⟨1287, 1319; 5⟩	⟨1307, 1302⟩	⟨**24**, **66**; **5**⟩	⟨**33**, **38**⟩	1456
	⟨1929, 1954; 6⟩	⟨1940, 1944⟩	⟨**24**, **66**; **5**⟩	⟨**59**, **55**⟩	1444
	⟨1931, 1962; 6⟩	⟨1940, 1945⟩	⟨**24**, **66**; **5**⟩	⟨**59**, **54**⟩	1431
	⟨1287, 1319; 5⟩	⟨1298, 1302⟩	⟨**24**, **66**; **5**⟩	⟨**59**, **55**⟩	1417
	⟨1183, 1221; 5⟩	⟨1210, 1206⟩	⟨1331, 1361; 7⟩	⟨1339, 1343⟩	1396
	⟨1046, 1089; 5⟩	⟨1085, 1081⟩	⟨1331, 1361; 7⟩	⟨1339, 1343⟩	1389
	⟨1183, 1221; 5⟩	⟨1199, 1203⟩	⟨1331, 1361; 7⟩	⟨1350, 1346⟩	1386
	⟨1716, 1758; 5⟩	⟨1742, 1738⟩	⟨1263, 1288; 6⟩	⟨1272, 1276⟩	1374
	⟨1715, 1734; 6⟩	⟨1722, 1726⟩	⟨1331, 1361; 7⟩	⟨1353, 1349⟩	1372
	⟨368, 400; 5⟩	⟨382, 386⟩	⟨1331, 1361; 7⟩	⟨1352, 1348⟩	1371
	⟨2139, 2180; 6⟩	⟨2150, 2154⟩	⟨**24**, **66**; **5**⟩	⟨**56**, **52**⟩	1370
	⟨1716, 1758; 5⟩	⟨1722, 1726⟩	⟨1331, 1361; 7⟩	⟨1353, 1349⟩	1367
	⟨580, 606; 5⟩	⟨586, 590⟩	⟨1263, 1288; 6⟩	⟨1279, 1275⟩	1365
	⟨1716, 1758; 5⟩	⟨1731, 1735⟩	⟨1263, 1288; 6⟩	⟨1279, 1275⟩	1353
PB1–PA	⟨**24**, **66**; **5**⟩	⟨**31**, **35**⟩	⟨**2062**, **2103**; **6**⟩	⟨**2097**, **2093**⟩	1406
	⟨**24**, **66**; **5**⟩	⟨**30**, **35**⟩	⟨**2062**, **2103**; **6**⟩	⟨**2098**, **2093**⟩	1400
	⟨1331, 1361; 7⟩	⟨1355, 1351⟩	⟨**2062**, **2103**; **6**⟩	⟨**2075**, **2079**⟩	1367
	⟨1263, 1288; 6⟩	⟨1270, 1274⟩	⟨**2062**, **2103**; **6**⟩	⟨**2097**, **2093**⟩	1377
PA–HA	⟨**2062**, **2103**; **6**⟩	⟨**2097**, **2093**⟩	⟨650, 681; 5⟩	⟨661, 665⟩	1375
HA–NP	⟨650, 681; 5⟩	⟨676, 672⟩	⟨454, 495; 5⟩	⟨461, 465⟩	1367
NP–NA	⟨454, 495; 5⟩	⟨473, 477⟩	⟨378, 423; 5⟩	⟨419, 415⟩	1439
	⟨454, 495; 5⟩	⟨461, 465⟩	⟨378, 423; 5⟩	⟨410, 406⟩	1413
NA–MP	⟨378, 423; 5⟩	⟨398, 402⟩	⟨735, 765; 5⟩	⟨761, 757⟩	1353
MP–NS	⟨735, 765; 5⟩	⟨760, 755⟩	⟨**38**, **81**; **5**⟩	⟨**50**, **55**⟩	1383
	⟨735, 765; 5⟩	⟨759, 755⟩	⟨**38**, **81**; **5**⟩	⟨**51**, **55**⟩	1379
	⟨735, 765; 5⟩	⟨760, 755⟩	⟨**38**, **81**; **5**⟩	⟨**50**, **55**⟩	1383

* **太字**はパッケージングシグナル位置を含む対合ヘアピンループ

表 2 は, リバースジェネティ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. ここで列項目の NCBI の欄は, 本論文で用いた NCBI データにおけるパッケージングシグナルの位置を示す.

表 3 は分節間対合ヘアピンループの 7 分節中 4 分節以上パッケージングシグナルと同じ位置に対合ヘアピンループがあり, $780 = 1560/2$(すなわち 50%) 以上の頻度で見つかった対合ヘアピンループの数を示す

表 3 より, PB2, PB1 は PA, HA, NA, NS と共にパッケージングシグナル位置を含む 8 分節間対合ヘアピンループが 50%以上で見つかる. しかし, PB2, PB1 が同時にパッケージングシグナル位置を含む 8 分節間対合ヘアピンループは 50%以上の頻度では見つからない. 特に, RNA 配列 NP はパッケージングシグナル位置を含む 8 分節間対合ヘアピンループが形成されるのを妨げていると考えられる.

5 まとめ

本論文では, 配列間のミスマッチのない**対合ヘアピンループ**を導入した. また, 簡易化した Nussinov アルゴリズムによってヘアピンループをすべて抽出するアルゴリズム Hairpin, 2 配列間の極大対合ヘアピンループをすべて抽出するアルゴリズム Kissing, 8 分節間対合ヘアピンループを抽出するアルゴリズム Common を設計した. 最後に, いくつかの閾値を導入し, 8 分節間極大対合ヘアピンループを探索した. 特にパッケージ

表 2: RNA 分節のパッケージングシグナルの位置

RNA	NCBI
PB2	35-144, 2209-2304
PB1	38-163, 2197-2299
PA	38-124, 691-731, 742-767 2094-2156, 2169-2176
HA	38-125, 1659-1671
NP	46-165, 1482-1526
NA	35-185, 1211-1413
MP	ε^*
NS	890, 36-56

* ε は NCBI に該当番号が存在しないことを示す

表 3: 最大頻度 7 本中 4 本の RNA 分節におけるパッケージングシグナルと位置を共有する対合ヘアピンループ及び閾値 $(5, 5, 50, 5)$ の低い頻度での対合ヘアピンループの数

RNA segments	freq.	%	#kh
PB2, PA, HA, NA	931	59.68	2027
PB2, PA, HA, NS	973	62.37	90
PB2, PA, NA, NS	973	62.37	177
PB2, HA, NA, NS	973	62.37	65
PB2, PA, HA, NA, NS	907	58.14	65
PB1, PA, HA, NA	931	59.68	132
PB1, PA, HA, NS	1215	77.88	132
PB1, PA, NA, NS	1275	81.73	108
PB1, HA, NA, NS	907	58.14	151
PB1, PA, HA, NA, NS	907	58.14	135
NP, S $(S \subseteq \{\mathbf{PB2}, \mathbf{PB1}, \mathrm{PA}, \mathrm{HA}, \mathrm{NA}, \mathrm{NS}\})$	–	–	–
PB2, **PB1**	–	–	–

ングシグナル位置を含む極大対合ヘアピンループを探索した.

本論文ではヘアピンループと対合ヘアピンループにはミスマッチを含まないとしたが, ある程度のミスマッチを許容しつつ, すべてのヘアピン対合を求めるアルゴリズムの設計が今後の課題である.

アルゴリズム Hairpin に代わり, 線形時間で動く接尾辞木または接尾辞配列 [6] を基としたすべてのヘアピンループを抽出するアルゴリズムの設計は可能である. しかしこのようなアルゴリズムを設計することができたとしても, まだ定理 2 のアルゴリズム Kissing における計算時間量を減らすことができない. したがって, 接尾辞木をアルゴリズム Hairpin, Kissing に組み込むことによって時間計算量を低減し, 効率的なアルゴリズムを設計するのが今後の課題となる.

表 3 より, PB2, PB1, PA, HA, NP, NA, NS から順に対合ヘアピンループを収集する. RNA 分節におけるパッケージングシグナル位置を含む対合ヘアピンループをより詳細に解析するために, 分節間すべてのペアに対して比較する必要がある.

最後に, 抽出した対合ヘアピンループとパッケージングシグナルの関係をより詳細にウイルス学的視点 [4, 5, 9, 10] から, 分析することは今後の重要な課題となる. 特に, MP, NS の RNA 分節のパッケージングシグナル位置はウイルス学では明確になっていない. 本論文は, 対合ヘアピンループに関与しているパッケージングシグナルの新しい位置を提供することができる.

参考文献

[1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter; *Molecular biology of the cell* (5th ed.), Newton Press, 2008.

[2] Y. 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/`.

[3] R. Durbin, S. Eddy, A. Krogh, G. Mitchison: *Biological sequence analysis*, Cambridge University Press, 1998.

[4] K. Fujii, Y. Fujii, T. Noda, Y. Matsumoto, T. Watanabe, A. Takada, H. Goto, T. Harimoto, Y. Kawaoka: *Importance of both the coding and the segment-specific noncoding regions of the influenza A virus NS segment for its efficient incorporation into virions*, J. Virol. **79**, 3766–3774, 2005.

[5] C. Gavazzi, M. Yver, C. Isel, R. P. Smyth, M. Rosa-Calatrava, B. Lina, V. Moulés, R. Marquet: *A functional sequence-specific interaction between influenza A virus genomic RNA segments*, Proc. Natl. Acad. Sci. U.S.A. **110**, 16604–16609, 2013.

[6] D. Gusfield: *Algorithms on strings, trees, and sequences*, Cambridge University Press, 1997.

[7] I. Hamada, T. Shimada, D. Nakata, K. Hirata, T. Kuboyama: *Classifying nucleotide sequences and their positions of influenza A viruses through several kernels*, Proc. ICPRAM'15, 342–347, 2015.

[8] D. K. Hendrix, S. E. Brenner, S. R. Holbrook: *RNA structural motifs: Building blocks of a modular biomolecule*, Quart. Rev. Biophysics **38**, 221–243, 2005.

[9] 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.

[10] M. Ozawa, J. Maeda, K. Iwatsuki-Horimoto, S. Watanabe, H. Goto, T. Horimoto, Y. Kawaoka: *Nucleotide sequence requirements at the 5' end of the influenza A virus M RNA segment for efficient virus replication*, J. Virol. **83**, 3384–3388, 2009.

[11] T. Shimada, T. Hazemoto, S. Makino, K. Hirata, K. Ito: *Finding correlated mutations among RNA segments in H3N2 influenza viruses*, Proc. SCIS-ISIS'12, 1702–1707, 2012.

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