

高機能スーパーコンピュータで可能になる生命科学分野での課題

全ゲノム・全タンパク質配列の 自己組織化マップを用いた大 規模ポストゲノム解析

プロジェクト責任者 池村淑道

プロジェクトメンバー 阿部貴志

長浜バイオ大学 バイオサイエンス学部

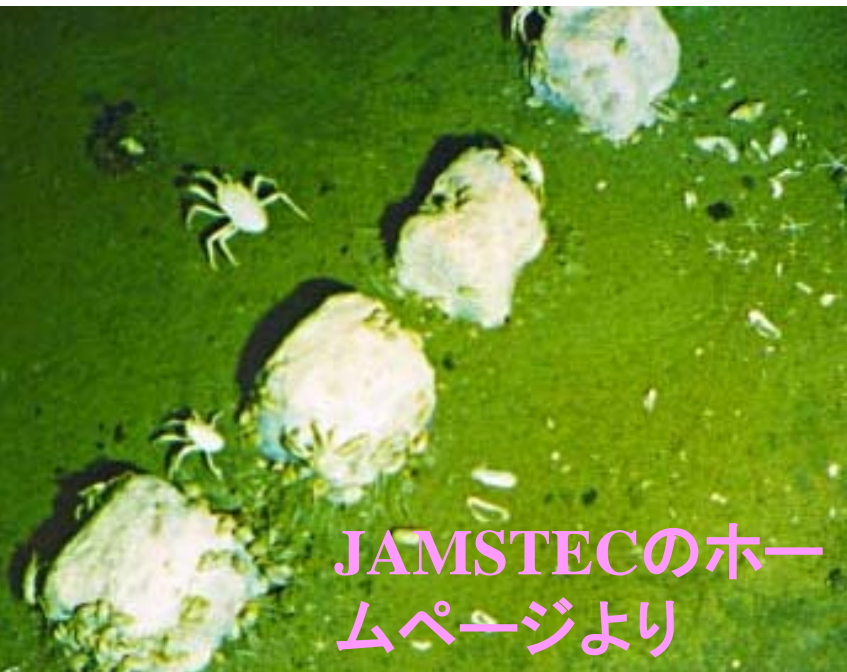
成果の産業利用を主たる視点とした。

完成した部分と応用分野の拡
張のための開発中の部分

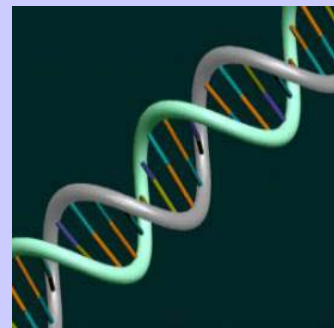
Metagenomes: メタゲノム解析

(膨大な未開拓ゲノム資源の活用)

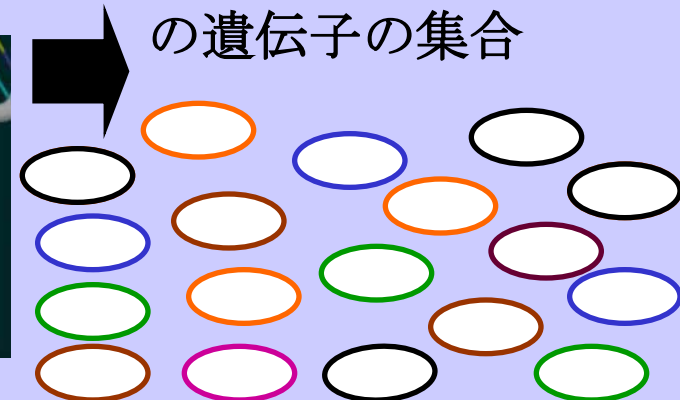
難培養性を含む微生物 のゲノムDNAを自然環境から微生物の培養・分離をすることなく、**混合ゲノム試料**として回収し、配列決定を行い、**産業的に有用な遺伝子**を探索する。



環境から抽出したDNAの断片化



混合ゲノム試料の遺伝子クローン集合：
どのゲノムに由来したのかが不明な多数の遺伝子の集合



自然環境に生息する99%以上の微生物は培養が困難で、未開拓なゲノム資源として残されて来た。最近、培養過程を経ずに大量なゲノム断片配列を解読する技術「メタゲノム解析法;環境ゲノム解析」が発展し、新規性の高い未知生物種のゲノムデータが加速度的に蓄積している。産業利用ならびに全地球レベルでの微生物群集の生態系の把握が可能になる。しかし、数千万件程度のゲノム断片配列が解読されているが、その大半の配列の由来する生物系統が不明である。

一括学習型自己組織化マップBLSOM

生命の設計図であるゲノムは、4種類の文字(A, T, G, C; 塩基と呼ぶ)で書かれている。

ACAGATTAGACCCTGAC-----

例えば、ヒトゲノムの場合は、30億文字(3Gb)で書かれており、朝刊の新聞に例えると、25年分。全体で2000年分。
ページ別にバラバラにして、生物種ごとに再集合可能？

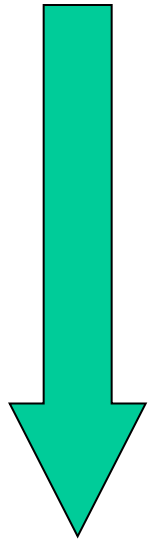
塩基配列が既知なすべての生物のゲノム配列を対象に、各々を1万文字(10 kb)に断片化して以下の単語を数える。

- 2連塩基: AA, AC, -----: **16種類**の単語
- 3連塩基: AAA, AAC, -----: **64種類**の単語
- 4連塩基: AAAA, AAAC, -----: **256種類**の単語
- 5連塩基: AAAAA, AAAAC, ---: **1024種類**の単語

高次元の大量情報解析のための**地球シミュレータ**の利用

一括学習型自己組織化マップ法: **BLSOM**

ゲノム情報の解析に用いるために、入力データの順序に依存しない形に変更: 大規模並列計算に適する



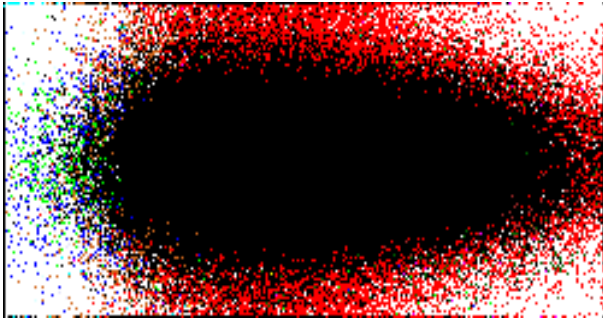
ゲノム配列が判明している13種の真核生物

10kb (一万文字分; 新聞1頁)及び
100kb (十万文字分)ごとに断片化して
3連と4連続文字頻度を計算

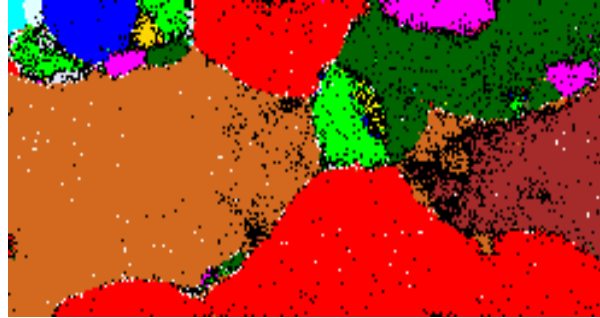
各々の生物種のゲノム配列に潜む生物種に固有な特徴を検出することができる。

真核生物13種のゲノム配列を対象にした 連続塩基の頻度に関するBLSOM解析

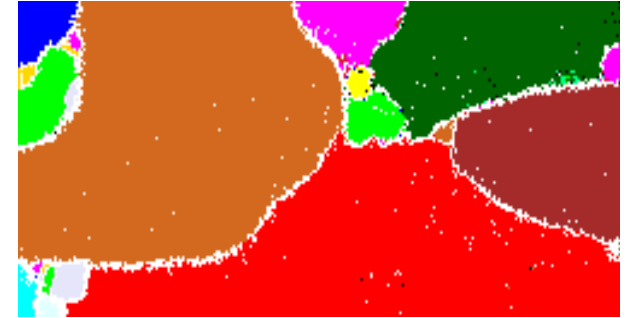
3連塩基PCA, 10-kb



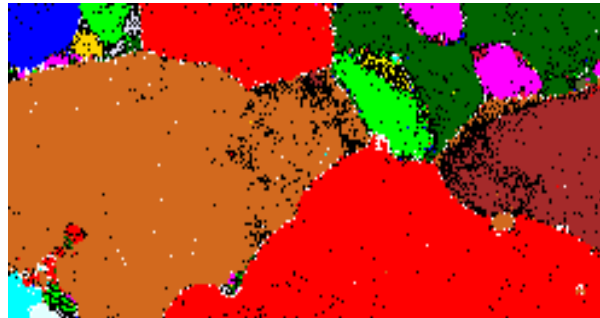
3連塩基SOM, 10-kb



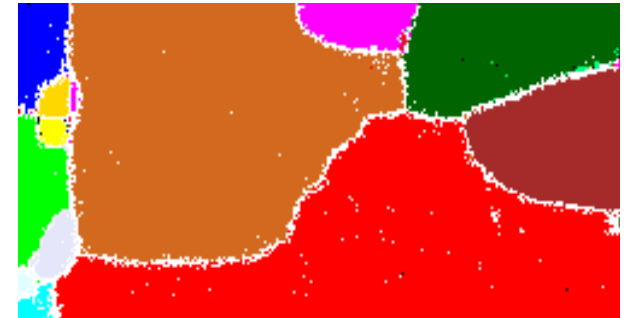
3連塩基SOM, 100-kb



4連塩基SOM, 10-kb



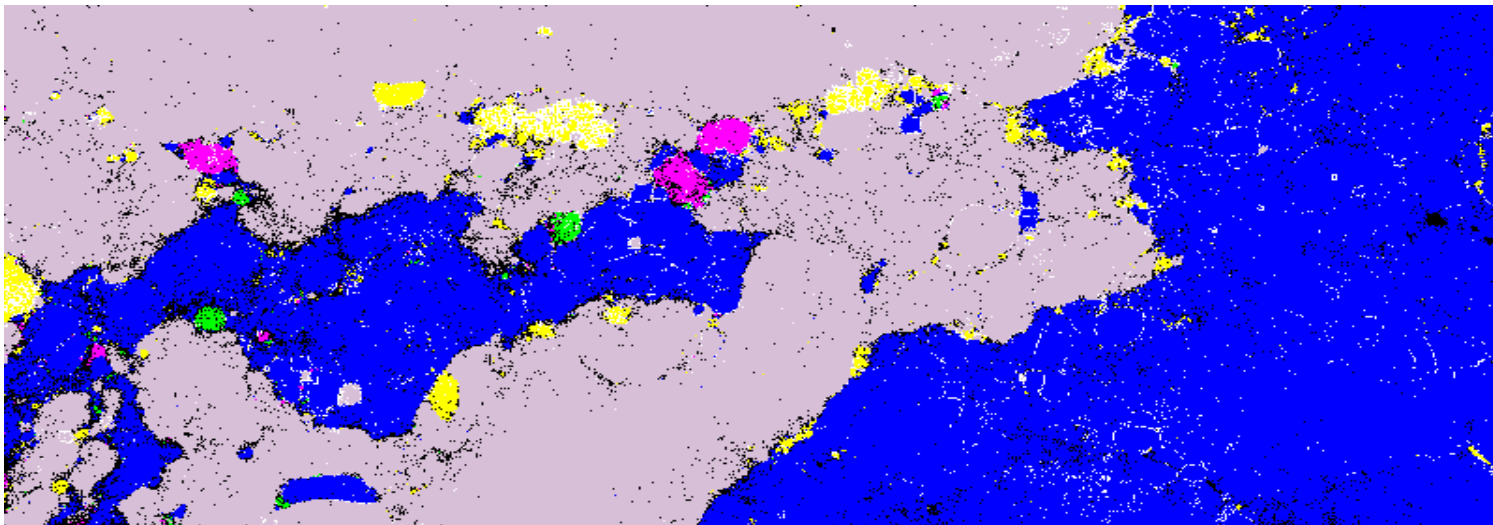
4連塩基SOM, 100-kb



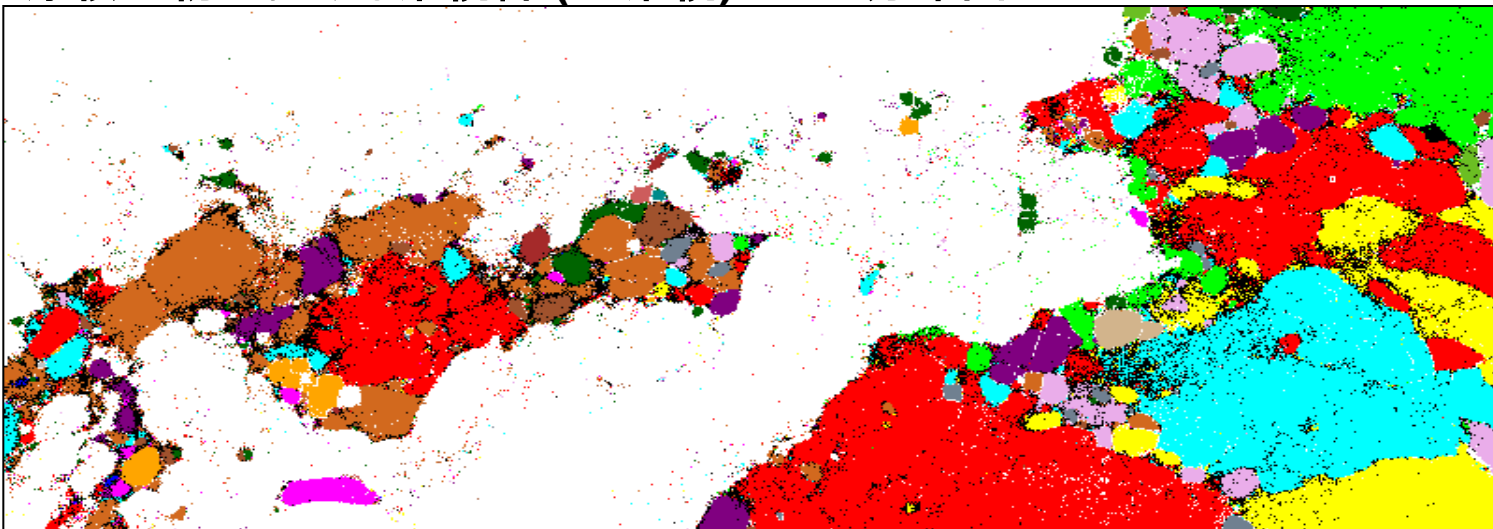
パン酵母 (■), 分裂酵母 (■),
粘菌(■),赤痢アメーバ(■),
マラリア原虫(■), シロイヌナズナ (■),
ウマゴヤシ (■), イネ (■), 線虫 (■),
ショウジョウバエ (■), フグ (■),
ゼブラフィッシュ (■), ヒト (■).

計算機には生物種名を与えずに、まず似た
単語の使い方をするものどうしを近づける。

原核生物2,813種、真核生物111種、ミトコンドリア1,728種、葉緑体 110種、ウイルス31,486種、断片化サイズ5 kb、縮退4連続塩基での大規模BLSOM



原核生物における系統群 (28系統)ごとの分布図



米国 NSFの「Cyber-enabled Discovery and Innovation Initiative」に関するMeetingの主催者側レポート中で、我々のグループがJournal of the Earth Simulatorで発表した図の一部をゲノム解析の例として掲載したいとの依頼があった。

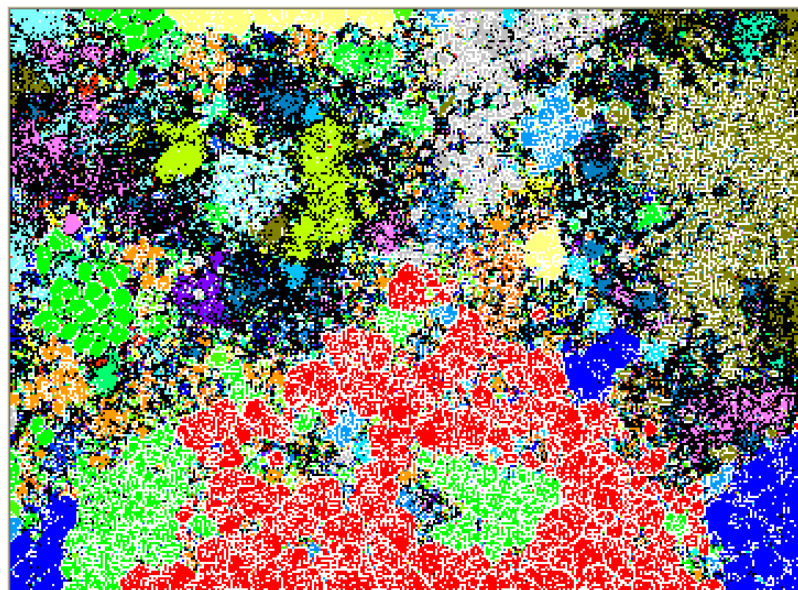
BLSOMでウイルスゲノム配列にも良い分離が見られたので、成果の産業や医学への早急な応用を目指して、予備的な研究開発を開始した。

来年度中に完成の予定

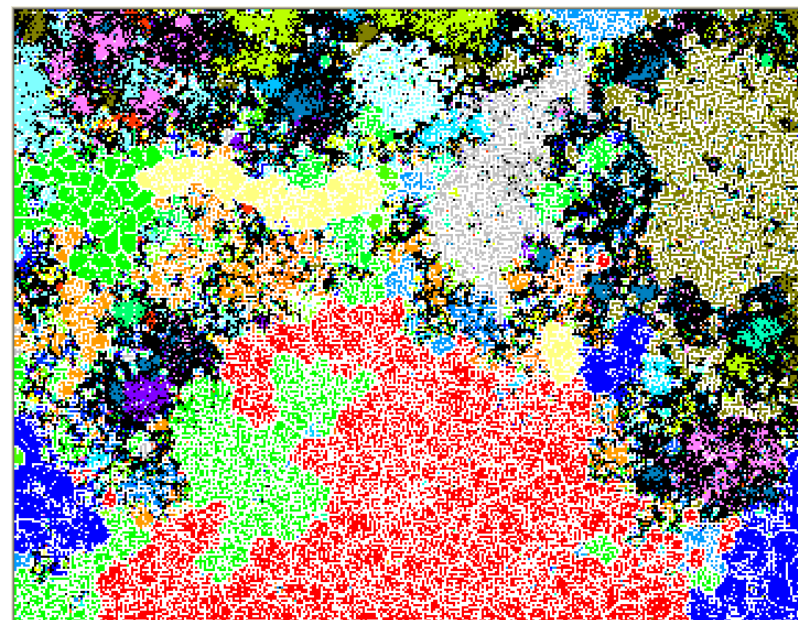
「全インフルエンザAウイルス株の俯瞰的可視化及びゲノム配列変化の予測」

全ウイルスを対象としたBLSOM解析(500b)

3連続塩基頻度



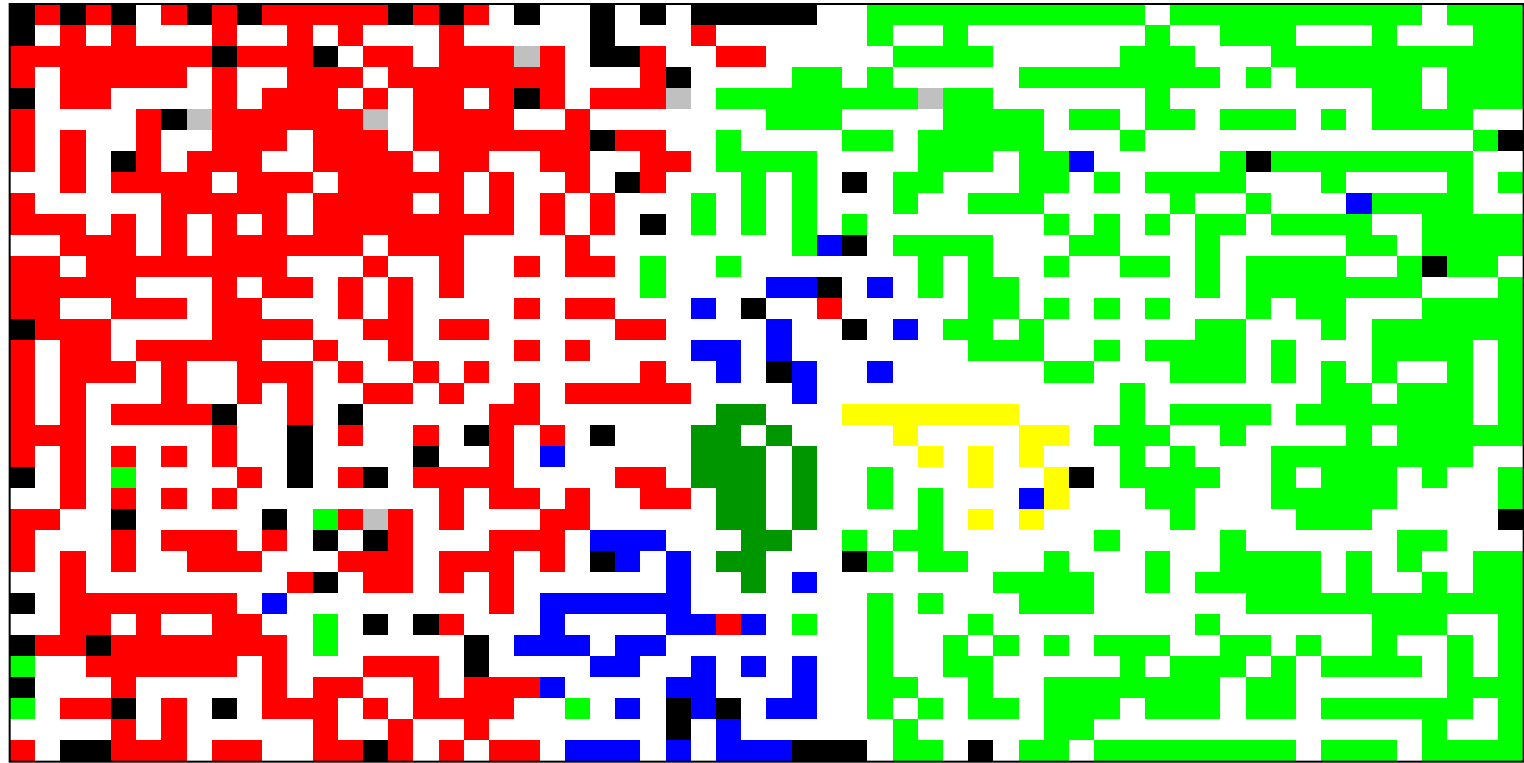
4連続塩基頻度



- | | | |
|--------------------|----------------|------------------|
| ■:Orthomyxoviridae | ■:Retroviridae | ■:Coronaviridae |
| ■:Herpesviridae | ■:Flaviviridae | ■:Poxviridae |
| ■:Hepadnaviridae | ■:Siphoviridae | ■:Picornaviridae |

DBに登録されている全ウイルスゲノム(43,828件)を500bごとに断片化した390,727配列の4連塩基頻度をBLSOM解析した。格子点が単一のウイルスfamilyの配列みで構成されていた場合には、各カテゴリーの色で着色し、複数のfamilyが混在している場合には黒にしている。

全インフルエンザAウイルス5350株を対象とした 4連続塩基頻度に基づいたBLSOM解析



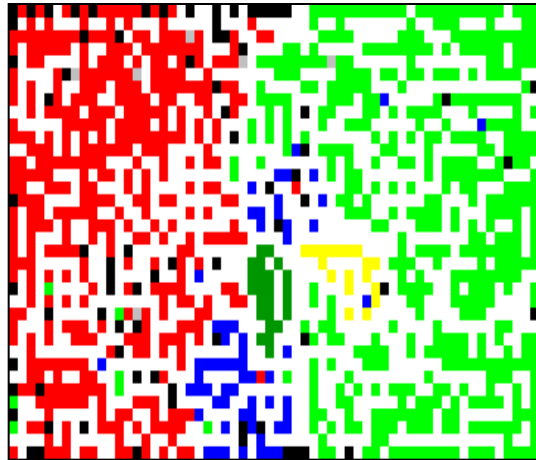
■ : Avian, 1948株 ■ : Human, 2955株 ■ : **新型** ■ : Equine, 68株
■ : Swine, 249株 ■ : Other (Seal, Tiger etc), 130株

単一の宿主生物に由来する配列のみが分離していた格子点は宿主カテゴリー別の色を着色し、複数の宿主由来配列が混在している場合には黒で示している。どの配列も分類されていない格子点は白色。

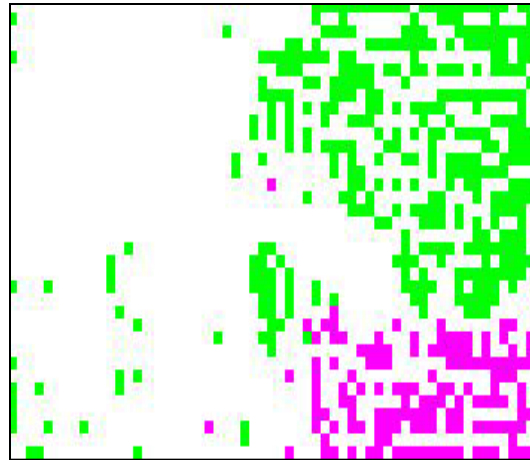
感染宿主ごとにウイルスゲノムの特徴が異なっていた。

ヒト由来ウイルスの亜型ごとの違い

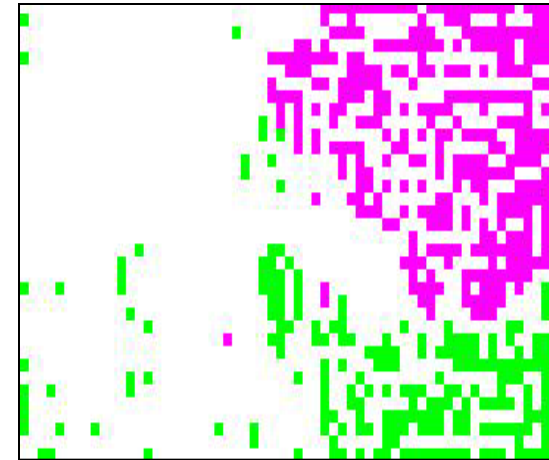
4連続塩基頻度



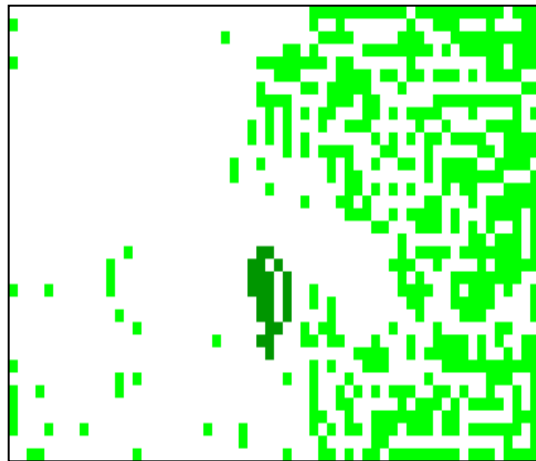
H1N1



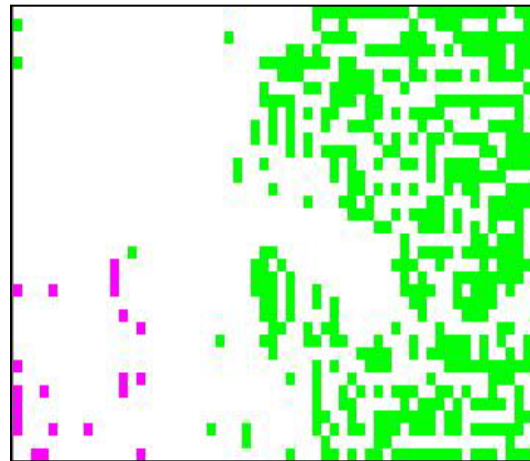
H3N2



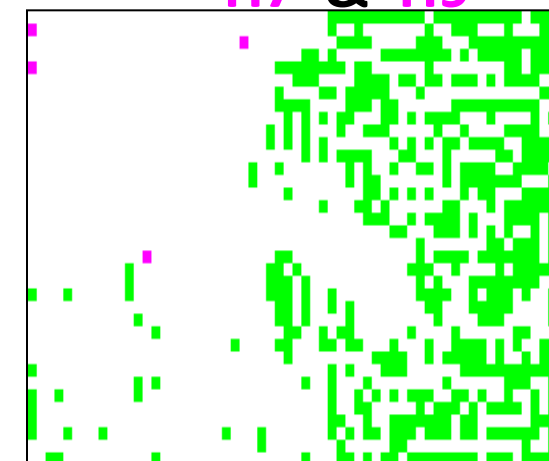
Human; 新型



H5N1



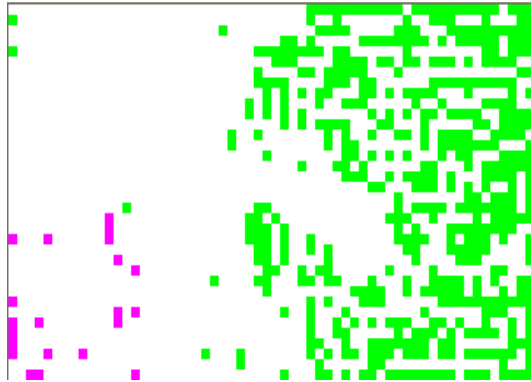
H7 & H9



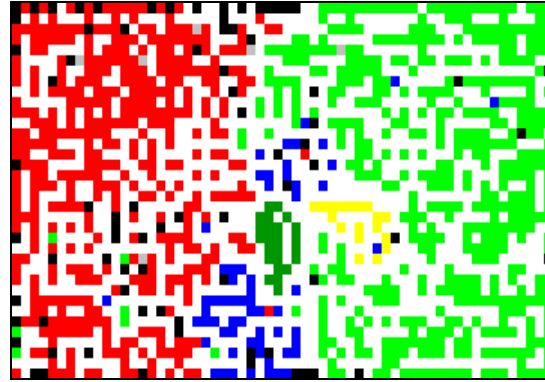
H5N1はヒト領域から離れた位置にある。新型H1N1もヒト領域から離れてはいるが、比較的近くに位置している。

次にヒトで流行を起こす可能性のあるトリの亜型や危険地域の予測

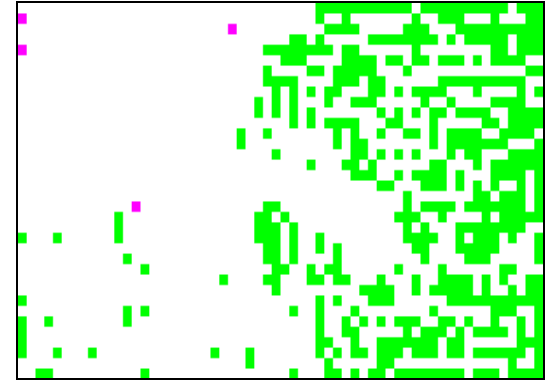
Human H5N1



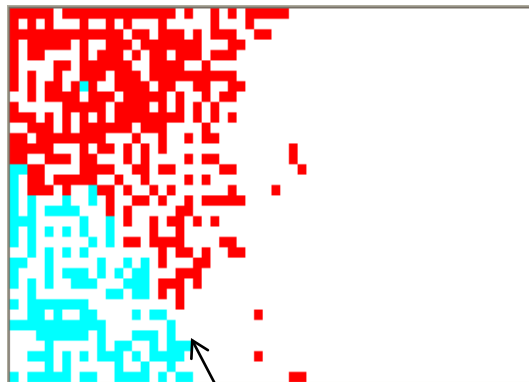
4連続塩基のBLSOM



Human H7N2 & H9N2

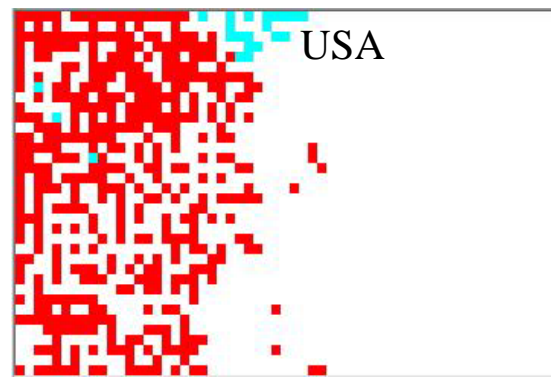


Avian H5N1



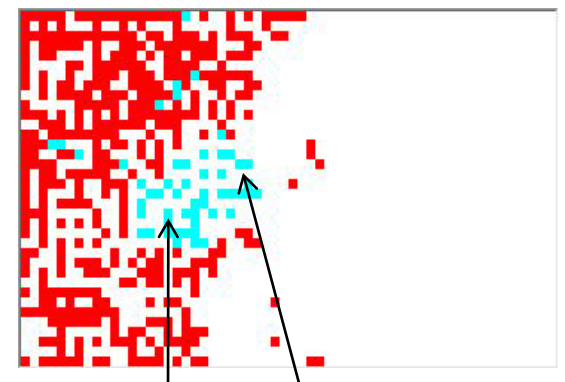
Russia
Kuwait
South Arabia

Avian H7N2



USA

Avian H9N2



China Israel

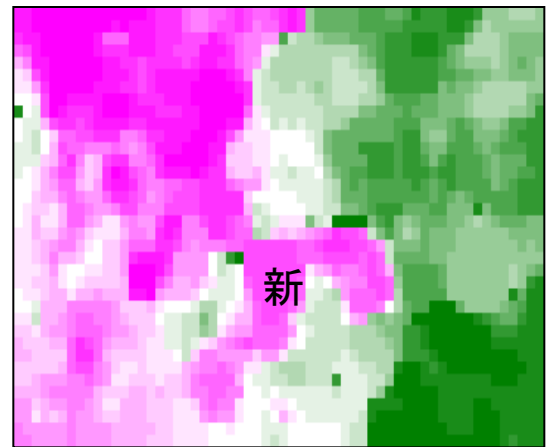
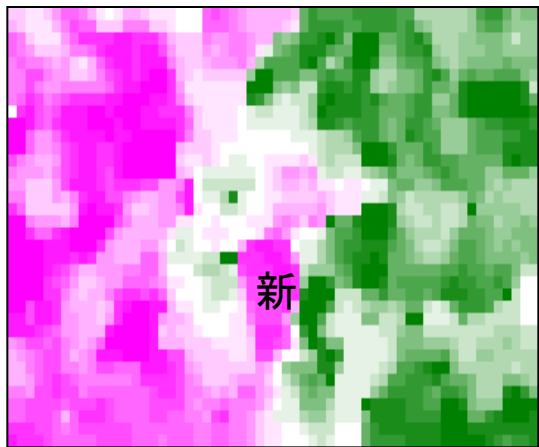
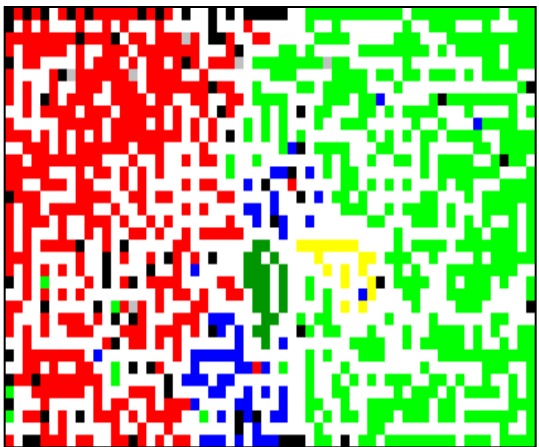
新型インフルエンザ株のオリゴヌクレオチド組成の一部は、季節性のヒト由来株からずれていて、トリ・豚・馬由来に近い。

高頻度:低頻度

4連続塩基のBLSOM

AGCG

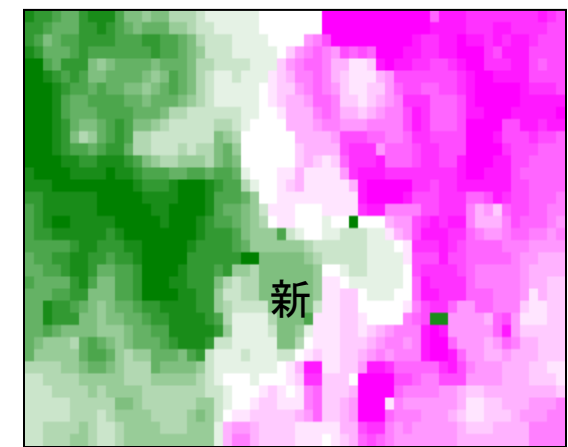
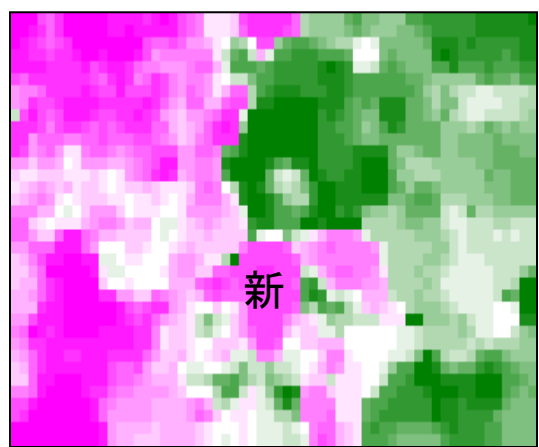
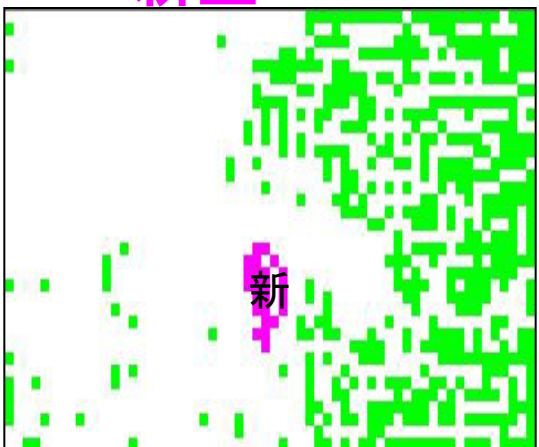
CCAC



新型H1N1

CGGC

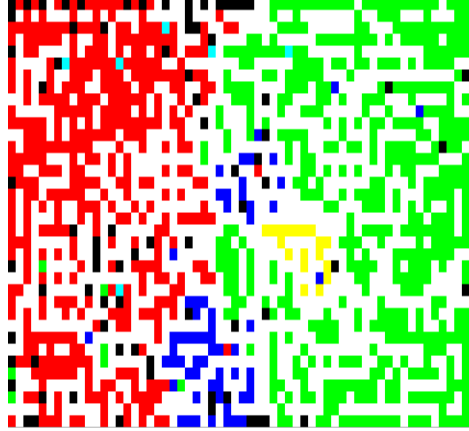
UUUU



これらのオリゴヌクレオチドは次第にヒト由来型に変わると予想してよいか？
そうならば、変化の方向を予測できる。一年後に検証可能

ヒト由来ウイルスの過去の時 系列的な変化

トリやブタの領域から発して遠ざかる傾向

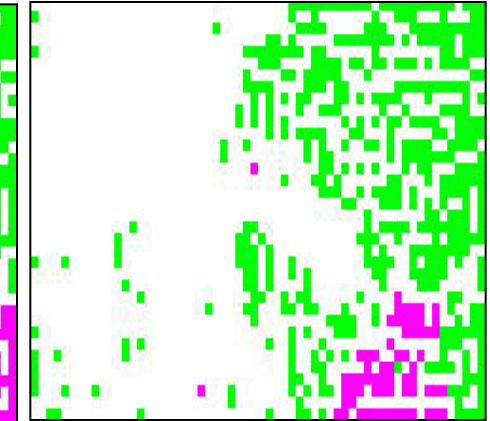
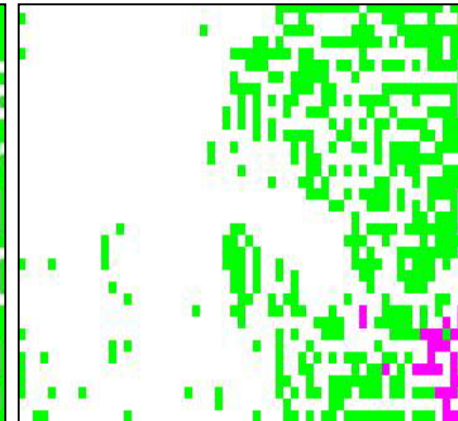
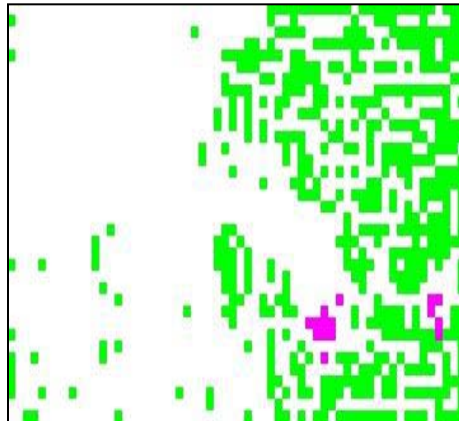
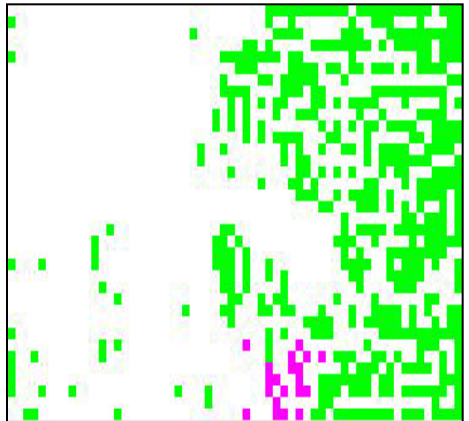


H1N1: 1977 - 1989

H1N1: 1990 - 1999

H1N1: 2000 - 2004

H1N1: 2005 - 2008

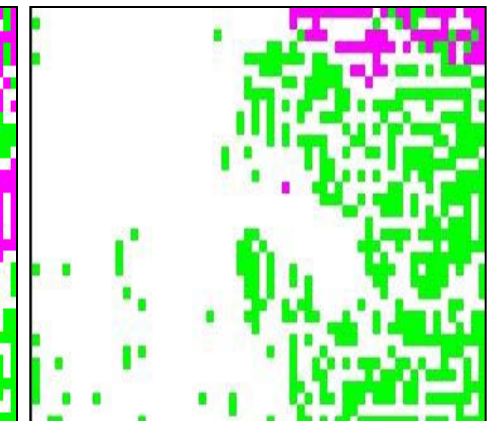
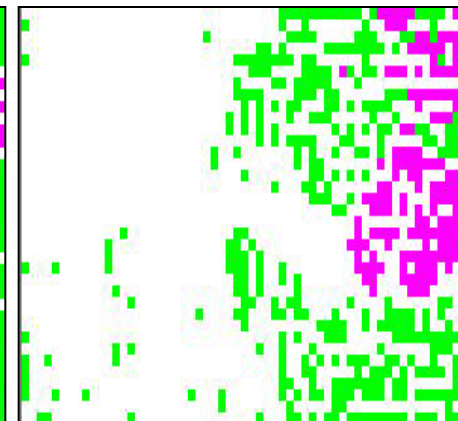
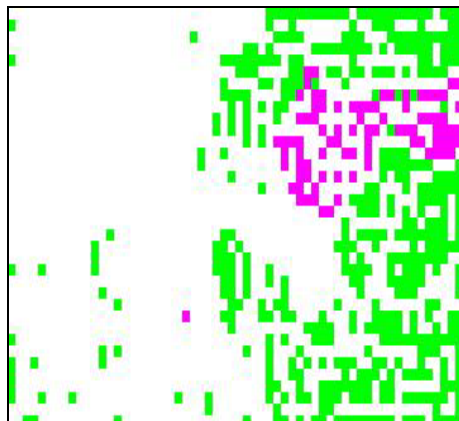
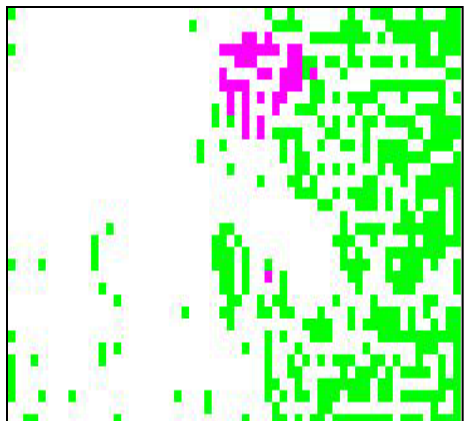


H3N2: 1968 - 1989

H3N2: 1990 - 1999

H3N2: 2000 - 2004

H3N2: 2005 - 2008



現時点での挑戦と計画

1. 新型H1N1が今後、どのように変化するか**の予測**。
2. 今後、パンデミックを引き起こす可能性のある亜型や株の**推定**、危険地域の**推定**。一年後に検証できる**有利さ**を最大限に活用したい。

来年度は、ESが可能にする、HIVを含む感染症対策のための大規模ゲノム解析法の確立を目指す。

DBに蓄積の著しい機能未知タン
パク質類の機能推定のための

BLSOMの開発->応用例の提示

配列相同性検索で機能が推定できない
タンパク質が2000万件程度蓄積している

が、それらの機能推定を行う方法の開発

が急務であり、産業界からの期待も大きい。

機能既知の全タンパク質の**BLSOM**

の年一回の更新。来年度は応用例の提示。