EARTH SIMULATOR

課題名：生物多様性を俯瞰するための大規模ゲノム情報基盤の整備

環境分野

課題代表者：阿部　貴志*1　課題参加者：池村　淑道*2

*1 新潟大学・工学部，*2 長浜バイオ大学

地球環境は多様な微生物類により多大な影響を受け、環境修復や保全における役割も大きい。近年のゲノム解読技術の発展は「メタゲノム解析」と呼ばれる新分野を生み、次世代シーケンサーの登場によって、全地球レベルでの生物生態系の把握を目標にした大規模解析が行われている。

ゲノム配列データの爆発的な増加に対応できる手法として、我々は高度な並列化に適したBLSOMを開発し、地球シミュレータを用いて、我が国のメタゲノム解析実験グループとの共同研究を継続している。

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。

現在は約4万種類のゲノムが解読されている。

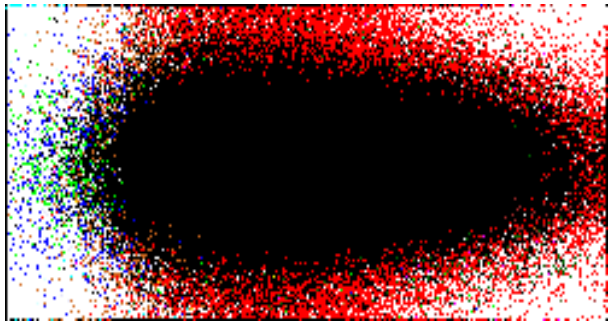
塩基配列が既知なすべての生物のゲノム配列を対象に、各々を1万文字(10 kb)に断片化して以下の単語を数える。

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

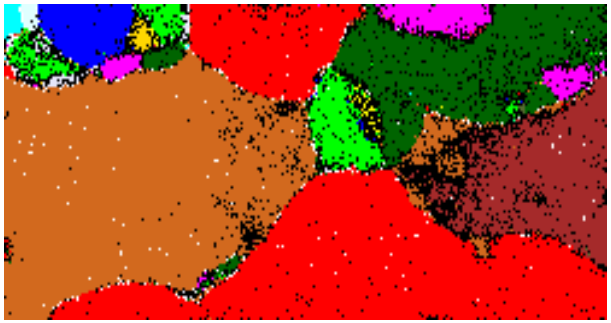
高次元の大量情報解析のため地球シミュレータを利用

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解析の例

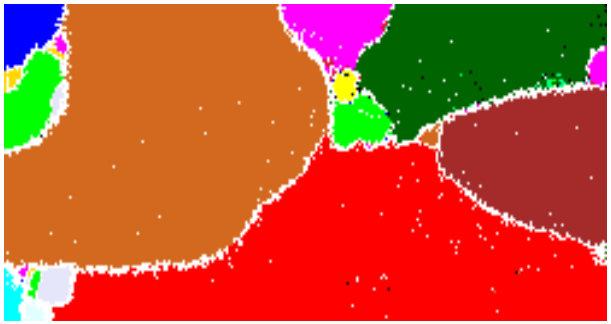
3連塩基PCA, 10-kb



3連塩基BLSOM, 10-kb

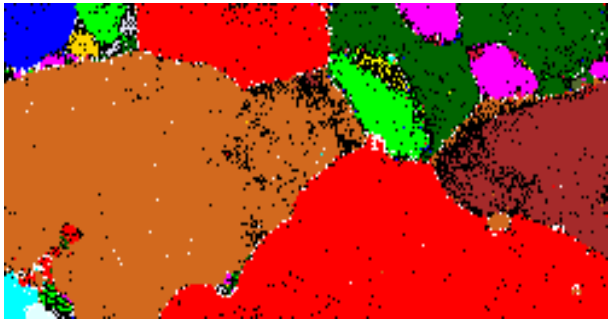


3連塩基BLSOM, 100-kb

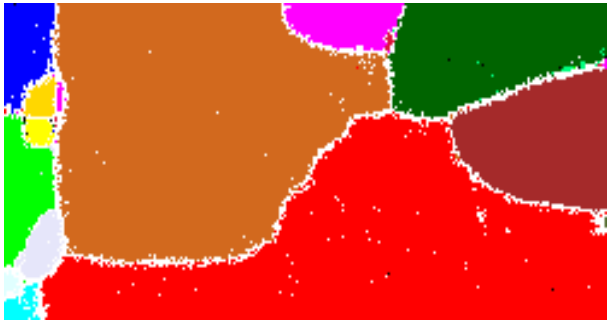


パン酵母（■）, 分裂酵母（■）,
粘菌(■),赤痢アメーバ(■),
マラリア原虫(■), シロイヌナズナ（■）,
ウマゴヤシ（■）, イネ（■）, 線虫（■）,
ショウジョウバエ（■）, フグ（■）,
ゼブラフィッシュ（■）, ヒト（■）.

4連塩基BLSOM, 10-kb



4連塩基BLSOM, 100-kb

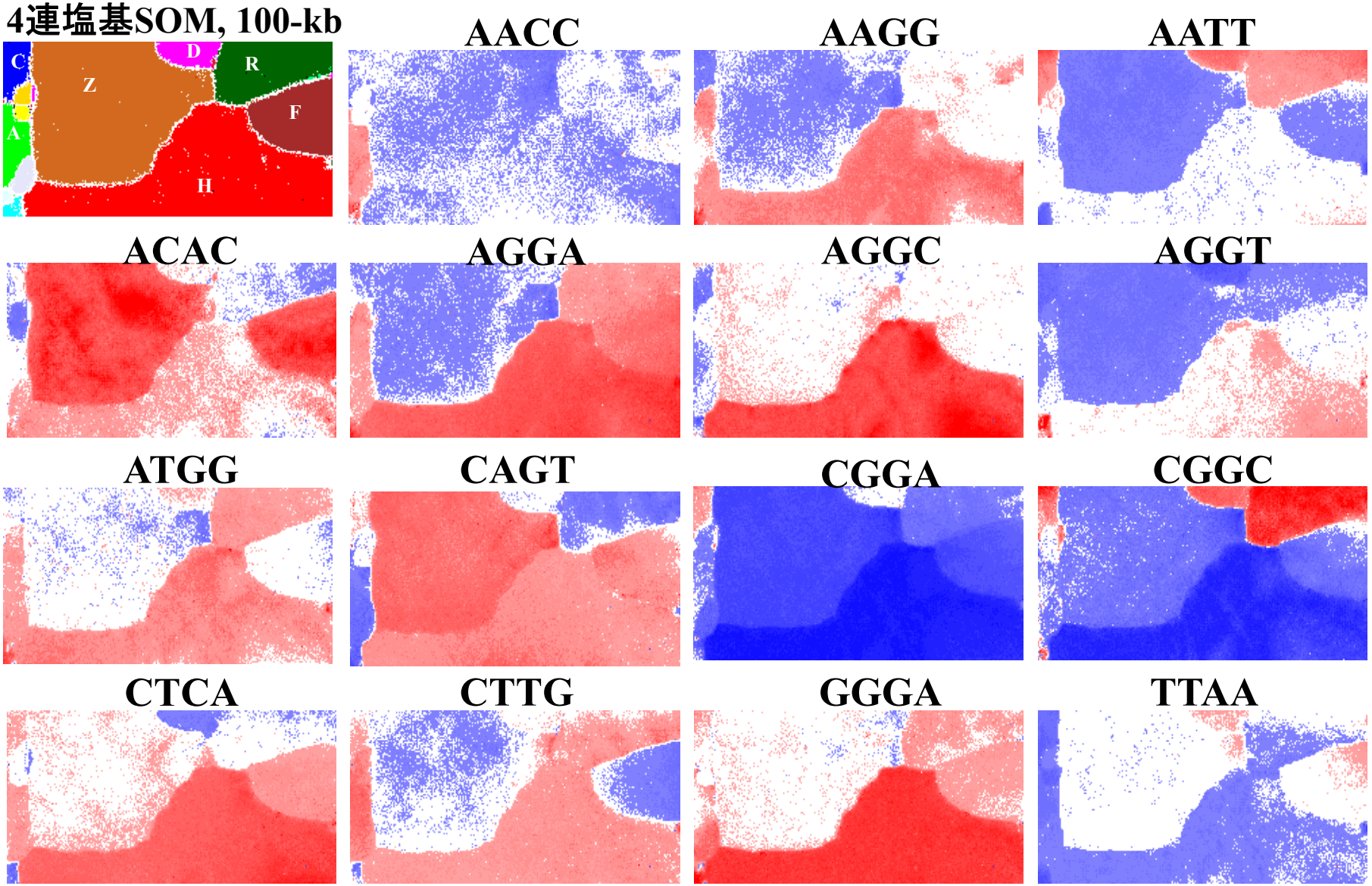


計算中に*生物種の情報を与えなくても*、生物種ごとに自己組織化。

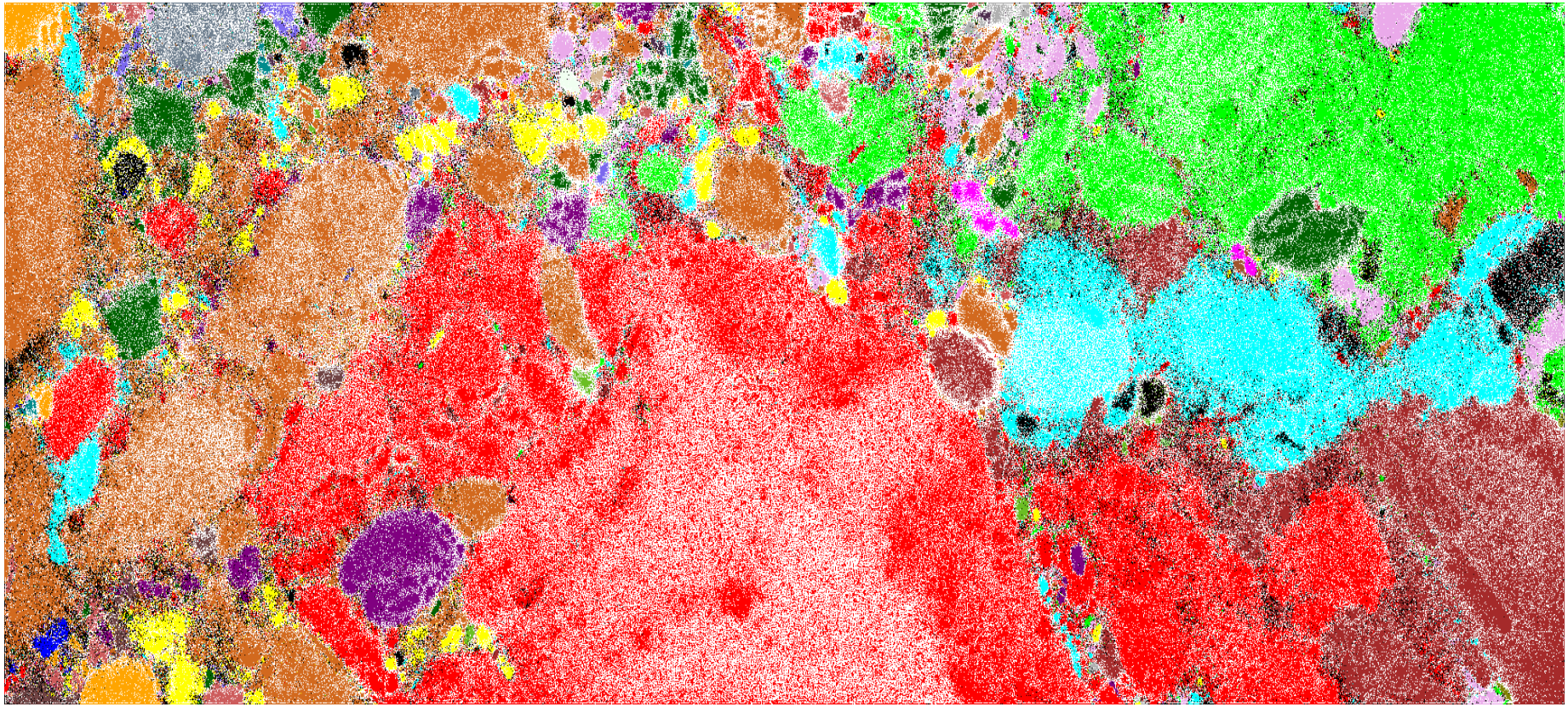
教師なしの機械学習。

Unsupervised data mining:予備知識やモデルや仮説なしに計算機が教えてくれる。

白, ランダム値; ■, 高頻度出現; ■, 低頻度出現



全既知原核生物3,457属のBLSOM（28の系統群に分離）



断片化サイズ5kb, 縮退4連続塩基でのBLSOMマップ
（解析データ数：3,868,729件，136次元ベクトル，2048コア利用）

強力な可視化機能により、全生物の多様性を俯瞰的に可視化
連続塩基出現頻度のみで原核生物および真核生物が高精度（97%）に分離
⇒連続塩基出現頻度の類似性のみで、生物系統が推定可能

メタゲノム解析①：一般的手法の現状

ウイルスは検出不可能

メタ16S rRNA解析

多様な環境から混合ゲノムDNAの抽出

16S rRNA 配列決定

相同性検索

既知微生物
16S rRNA 配列
データベース

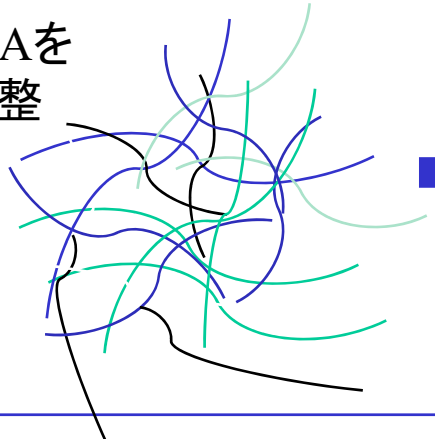
系統樹作成

微生物叢のゲノム
配列を丸ごと解読

遺伝子機能を知る

メタゲノム解析

全てのDNAを
抽出、調整



次世代シーケンサ

メタゲノム配列

AGTCTTAGCT

TTGAACCTA

CCGTCTTCTA

AATCCGGTG

メタゲノム解析②：一般的手法の問題点

新規性の高い生物由来ゲノム配列が多く存在

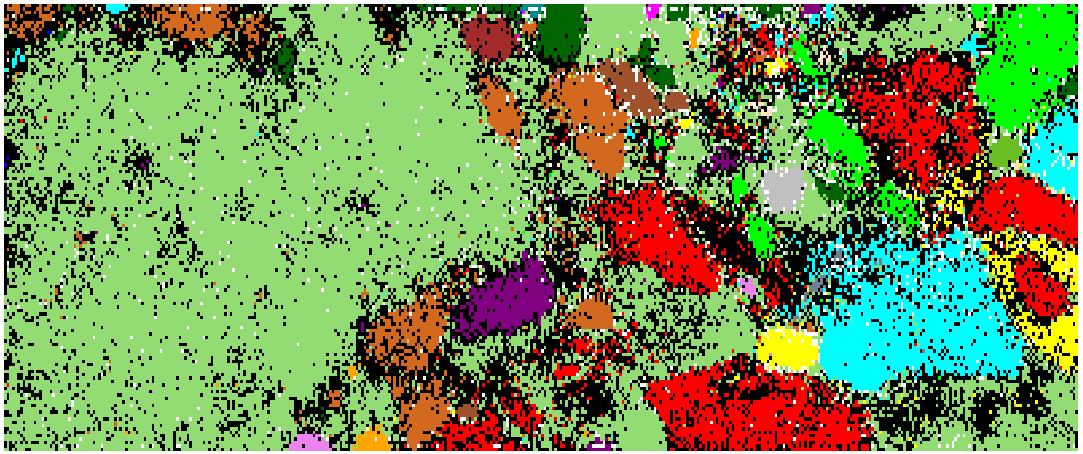
⇒既存の類似配列が少ないため，配列相同性検索では的確な生物系統推定が困難な場合が多い

配列相同性（配列間のアラインメント）とは
異なる観点での推定手法の開発が必須

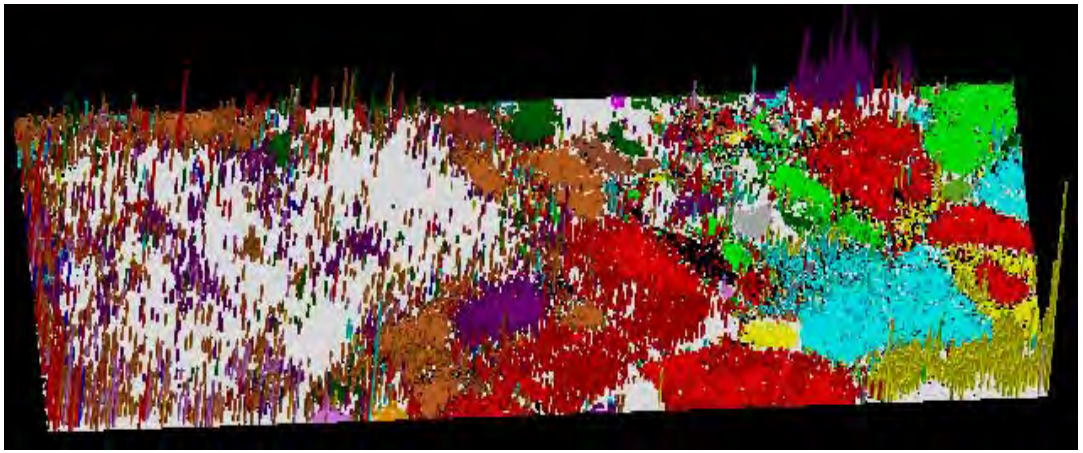
BLSOM

連続塩基組成に基づく一括学習型自己組織化マップ
（Batch Learning Self-Organizing Map; BLSOM）による
生物系統推定法を開発

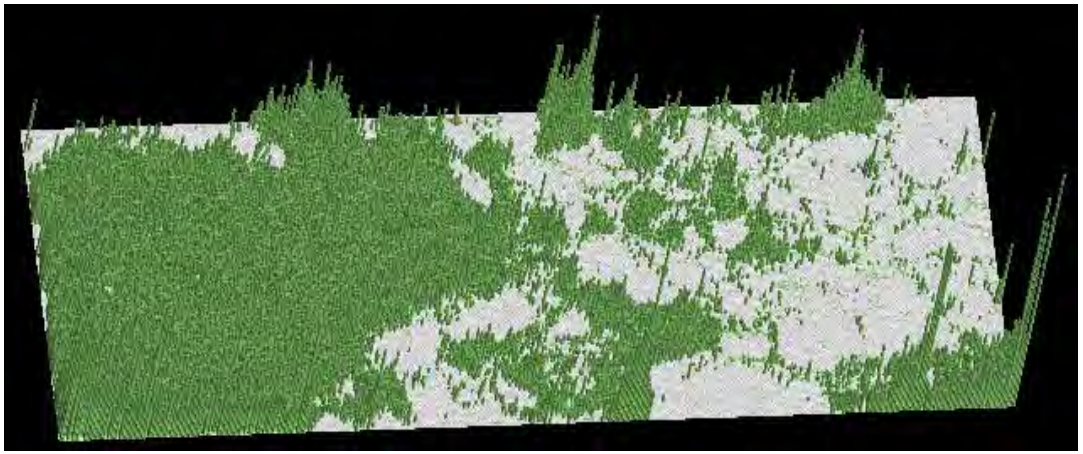
全既知原核生物
＋メタゲノム配列
(from Sargasso Sea)



メタゲノム配列で
既知生物と一緒に
クラスタリング（自己組織化）した(21%)



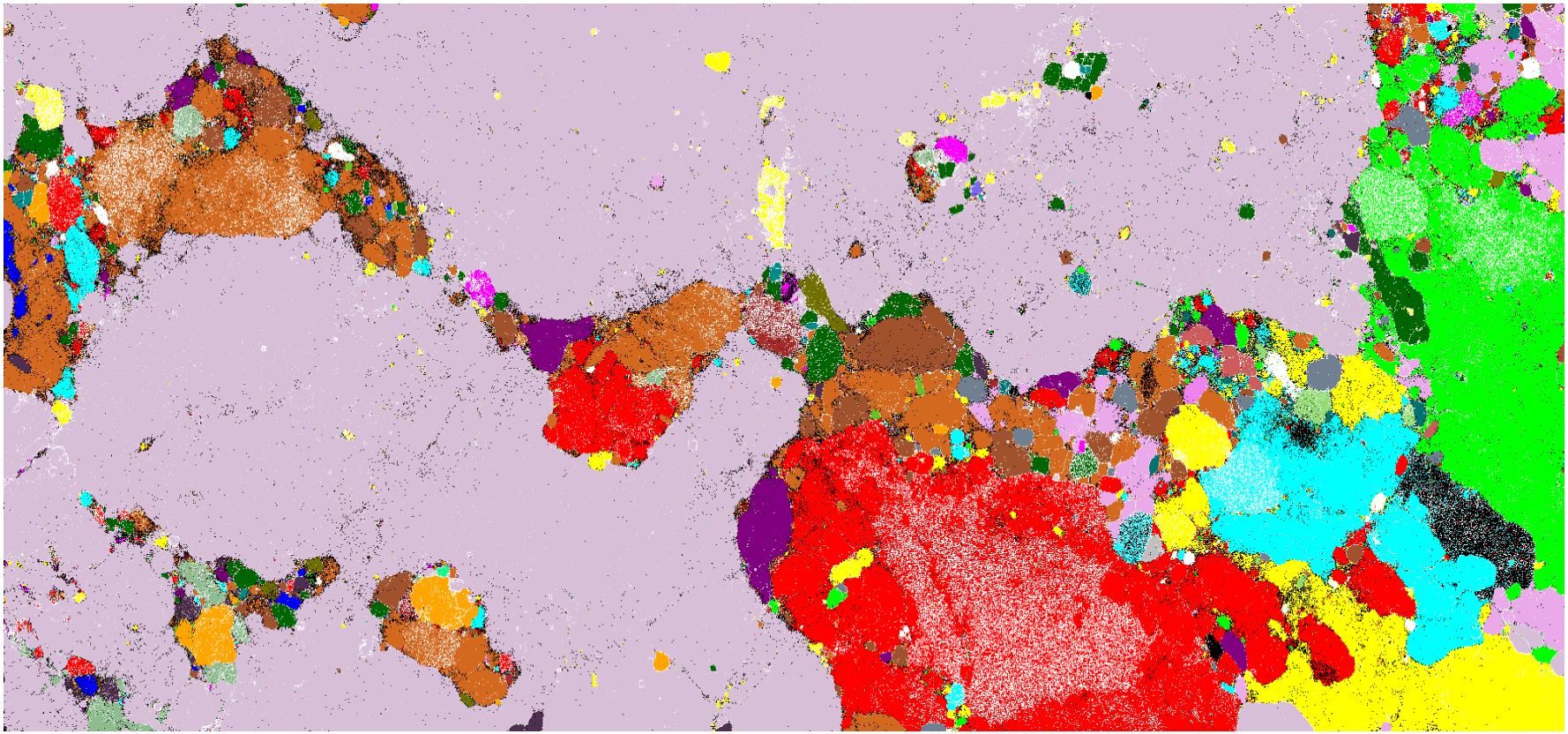
メタゲノム配列のみ
Unclassified (79%)
新規性の高いゲノム探索



最近は環境メタゲノム解析が注目されている

全既知生物種ゲノムのBLSOM　(毎年更新)

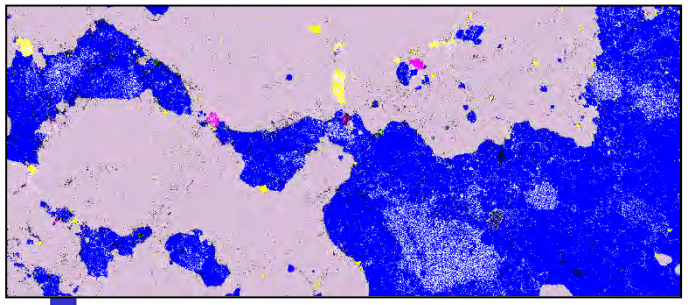
原核生物5,600種，真核生物412種，ミトコンドリア4,479種，
葉緑体 225種，ウィルス31,486種（1,120万件）での
断片化サイズ5 kb，縮退4連続塩基での大規模BLSOM



近年、ウィルスに対する要望が高まっている。

メタゲノム配列 (300塩基以上)

1st Step: Kingdom-BLSOM, 生物ドメインの推定

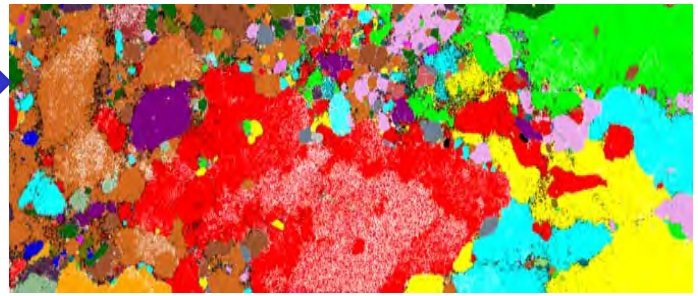


特徴

- 配列相同性とは異なるアプローチのため、相同性検索よりもロバスト性が高い
- 配列情報のみで、推定可能
- 段階的な予測により、新規性の高い微生物種の系統も検出可能

2nd Step: Prokaryote-BLSOM, 原核生物のPhylumの推定

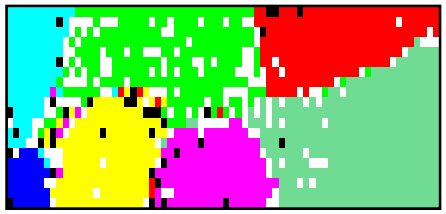
原核生物と推定された配列をマッピング



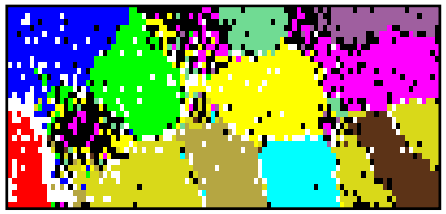
新型シーケンサなどにより大量のゲノム配列情報が次々と蓄積されており、各stepの参照用BLSOMマップを常に最新のものに更新する必要がある.

3rd Step: Genus-BLSOM, PhylumごとにGenusの推定

推定されたPhylumのBLSOMへマッピング



Actinobacteria



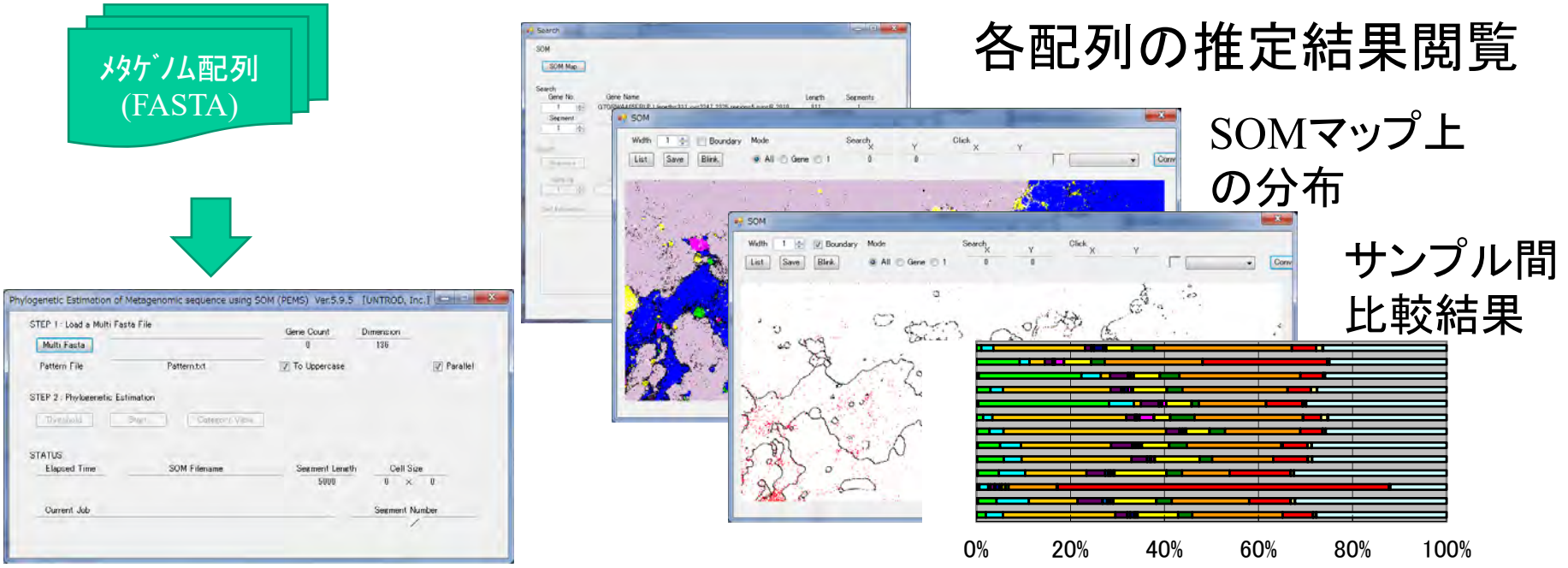
Alpha-proteobacteria

PEMS (Phylogenetic Estimation of Metagenomic sequence using BLSOM)

メタゲノム配列に対するBLSOMを用いた系統推定用ソフトウェアを公開

(ES成果の利用促進) 毎年更新したマップを使用

Kingdom ⇒ Phylum ⇒ Genusと多段階での予測が可能



多くの研究者に活用されている http://bioinfo.ie.niigata-u.ac.jp/?PEMS_Soft

巨大メモリーを備えたHPCが可能にするゲノム研究を目指して

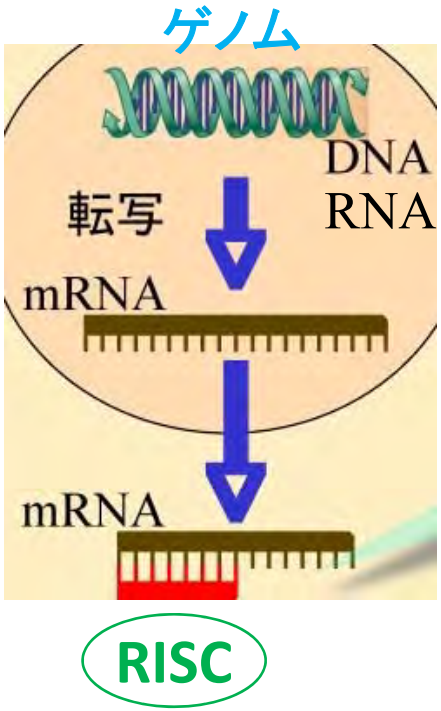
超高次元でスパースなビッグデータからの能率的な知識発見

探索的な試みの例

- 核酸医薬の英語名＝ therapeutic oligonucleotide
- 20～30連続塩基の核酸断片

20-mer（4^{20}) は約1兆1千億の変数
それなりの工夫が必要

- siRNA の場合
  RISCと呼ばれるたんぱく質複合体と結合してより効果的！

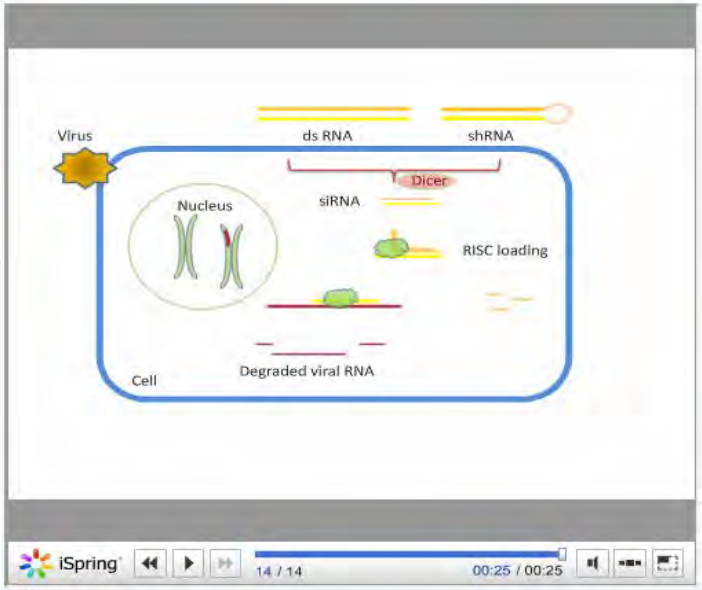


ウイルス用の核酸医薬のデータベース

20連程度の塩基のオリゴヌクレオチドを医薬品として使用

VIRsiRNAdb
Database of Viral siRNA / shRNA

Home Search Advance Search Browse EscapeDb Tools Submit About Team Contact



抗ウイルス核酸医薬用にデザインされ培養細胞系では活性が確認された例

- **VIRsiRNAdb** is a curated database of experimentally validated viral siRNA/shRNA targeting diverse genes of **42 important human viruses** including influenza, SARS and Hepatitis viruses.
- Currently database provides detailed experimental information of **1358 siRNA/shRNA** which includes siRNA sequence, virus subtype, target gene, GenBank accession, design algorithm, cell type, test object, test method and efficacy (mostly **quantitative efficacies**).
- Further, wherever available, information regarding **alternative efficacies** of above 300 siRNAs derived from different assays has also been incorporated.
- Database has facilities like search, advance search (using Boolean operators AND, OR) browsing (with data sorting option), internal linking and external linking to other databases (Pubmed, Genbank, ICTV).
- Additionally useful siRNA analysis tools are also provided e.g. **siTarAlign** for aligning the siRNA sequence with reference viral genomes or user defined sequences.
- VIRsiRNAdb would prove useful for RNAi researchers especially in siRNA based antiviral therapeutics development.

Reference:

Thakur Nishant, Qureshi Abid, Kumar Manoj*; VIRsiRNAdb: a curated database of experimentally validated viral siRNA/shRNA. Nucleic Acids Res. 2012 Jan;40(1):D230-6 10.1093/nar/gkr1147.

Related:

VIRsiRNApred:Viral siRNA prediction algorithm

インフルエンザA型ウイルスも重要なターゲット。 **45種類の配列が提案**

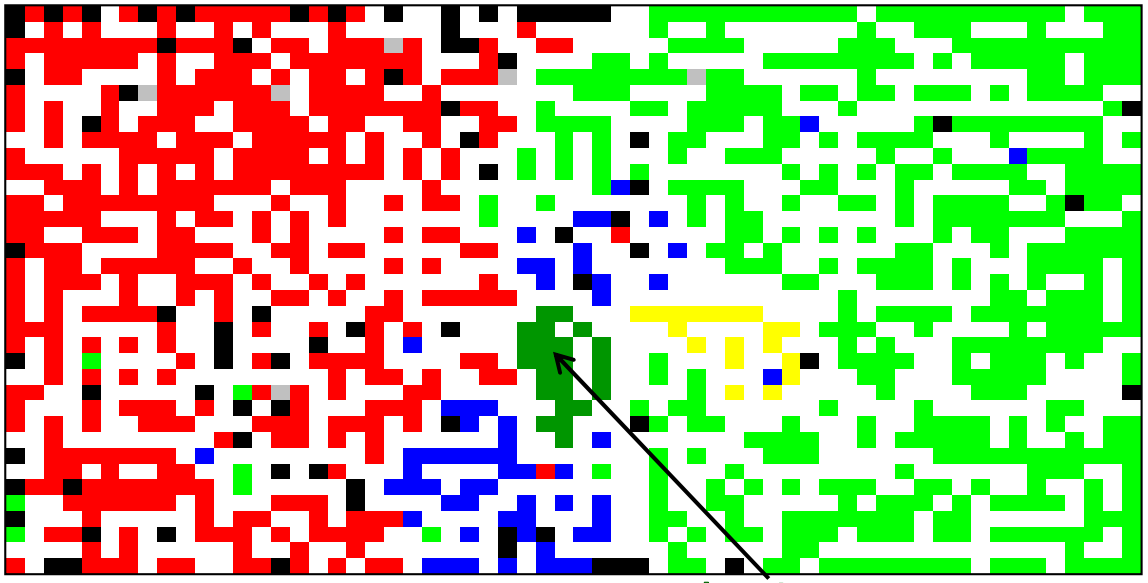
インフルエンザ、エボラ、マーズ、ジカ熱、デング熱のウイルスは極端に進化速度が高く、ある時期にデザインされた核酸医薬もその薬効を失いやすい。***待ち構え型の核酸医薬のデザイン。***

ウイルスに対して強力な手法を提供している。

インフルエンザ、エボラ、マーズウイルスの変化予測

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

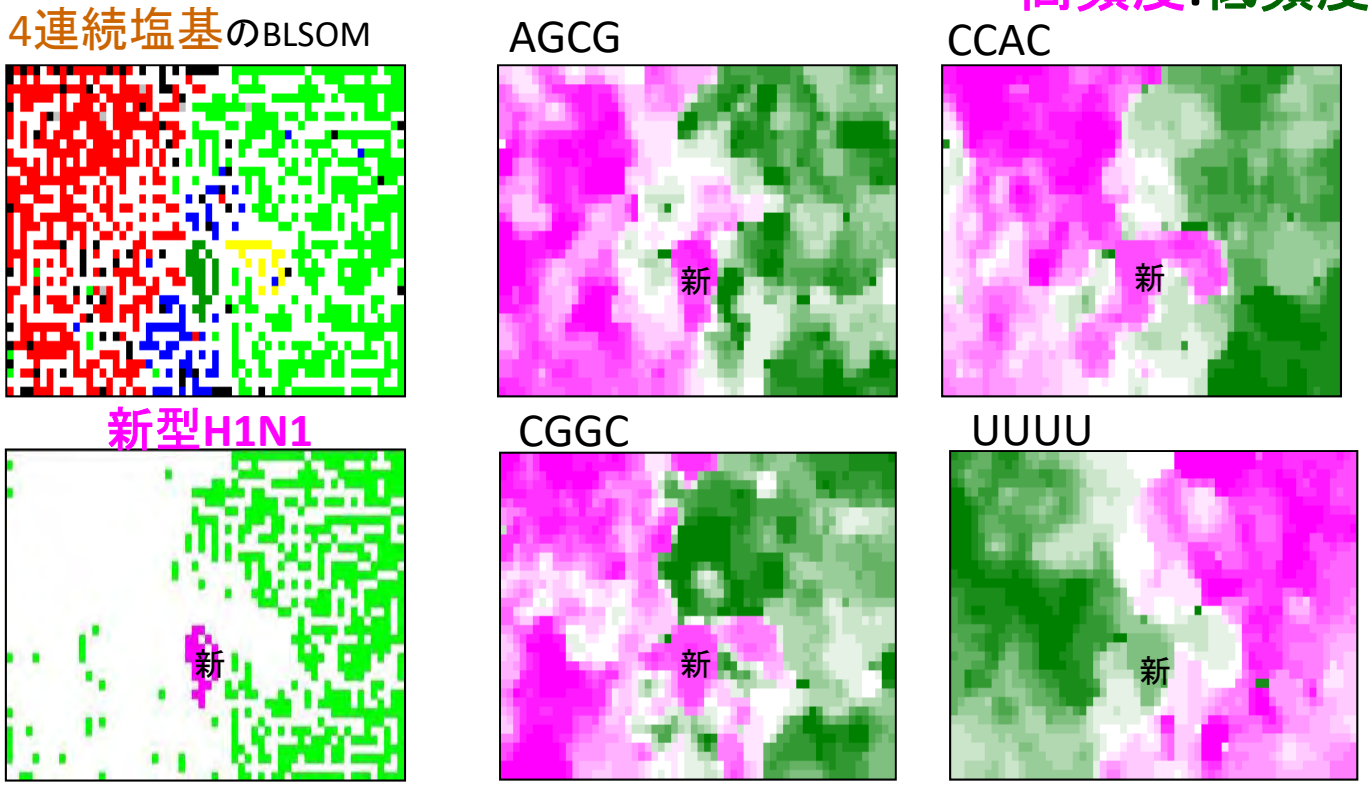
単一の宿主生物に由来する配列のみが分離していた格子点は宿主カテゴリー別の色を着色し、複数の宿主由来配列が混在している場合には黒で示している。どの配列も分類されていない格子点は白色。

*感染宿主ごとに*ウイルスゲノムの特徴が異なっていた。

BLSOM（教師無し機械学習）の発見

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



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

H1N1/09で変化が予想される連続塩基及びコドン

H1N1/09 では好まれているが、ヒト株では好まれない。 ***減ると予想***

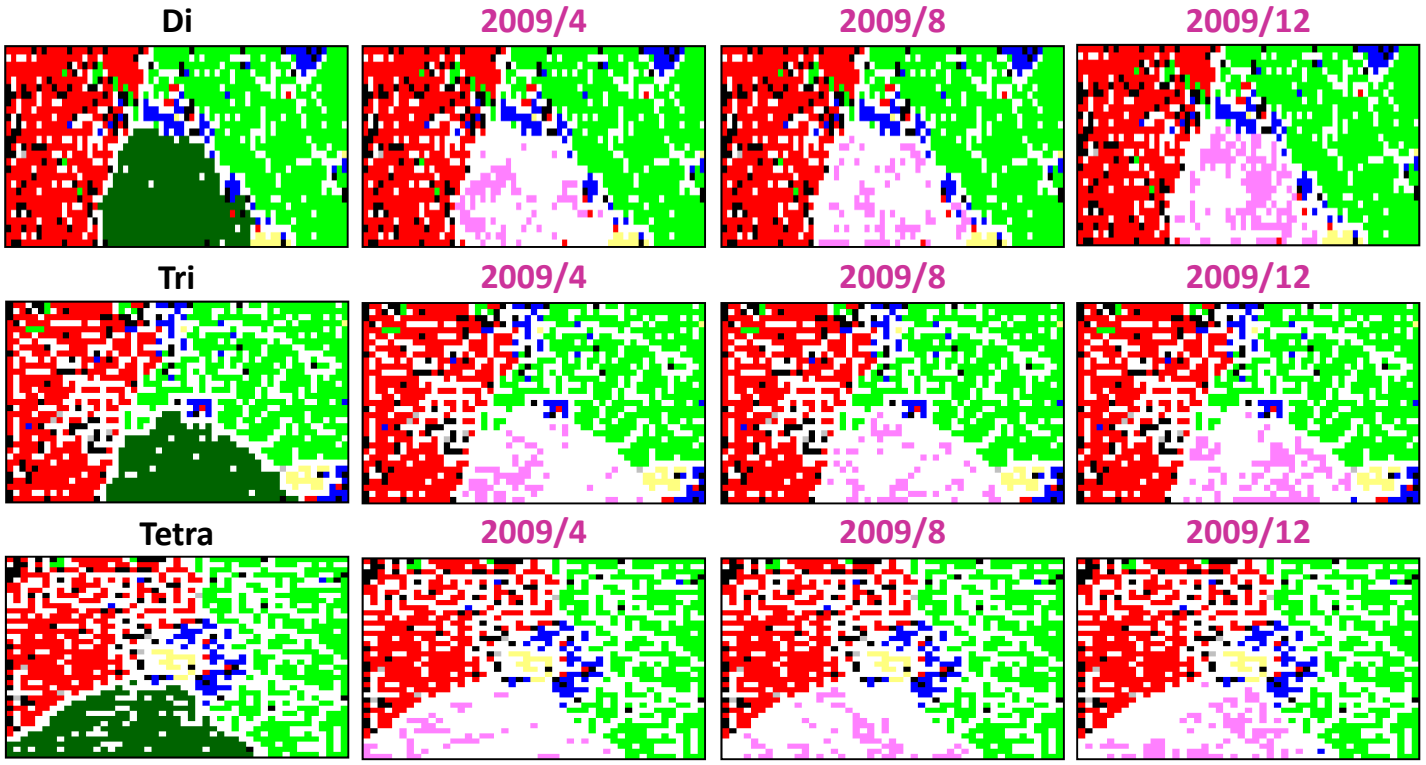
Codon	GCA, CAG, CUC, AAG, UUC, UCG
Di	AG, CG, GA
Tri	AGA, CAG, CCA, GCG, GUG
Tetra	AAGA, ACGG, AGAG, AGCG, AGGA, AUAA, AUCC, CACG, CCAC, CCAG, CGGC, GACG, GACU, GGCA, GUCG, GUCU, UCCA, UCUU, UGAA, UUCG

H1N1/09では好まれていないが、ヒト株で好まれる。 ***増えると予想***

Codon	CAA, UUG, AAA, UUU, ACU, GUU
Di	AA, UU
Tri	AAA, AUU, GGG, UCA, UGU, UUA, UUG, UUU
Tetra	AAAA, AAAC, AACU, AGCU, AUAG, AUUA, CAAA, GGGG, GGUU, GUCA, GUUG, UAUG, UGUA, UGUU, UUAA, UUAU, UUGU, UUUG, UUUU

2010年に予測を論文発表（DNA Res.　2011）した結果が、2012年に実証された（*BMC Infectious Diseases* ）

流行開始の半年後には、H1N1/09の2,143 株のゲノム配列が解読された。



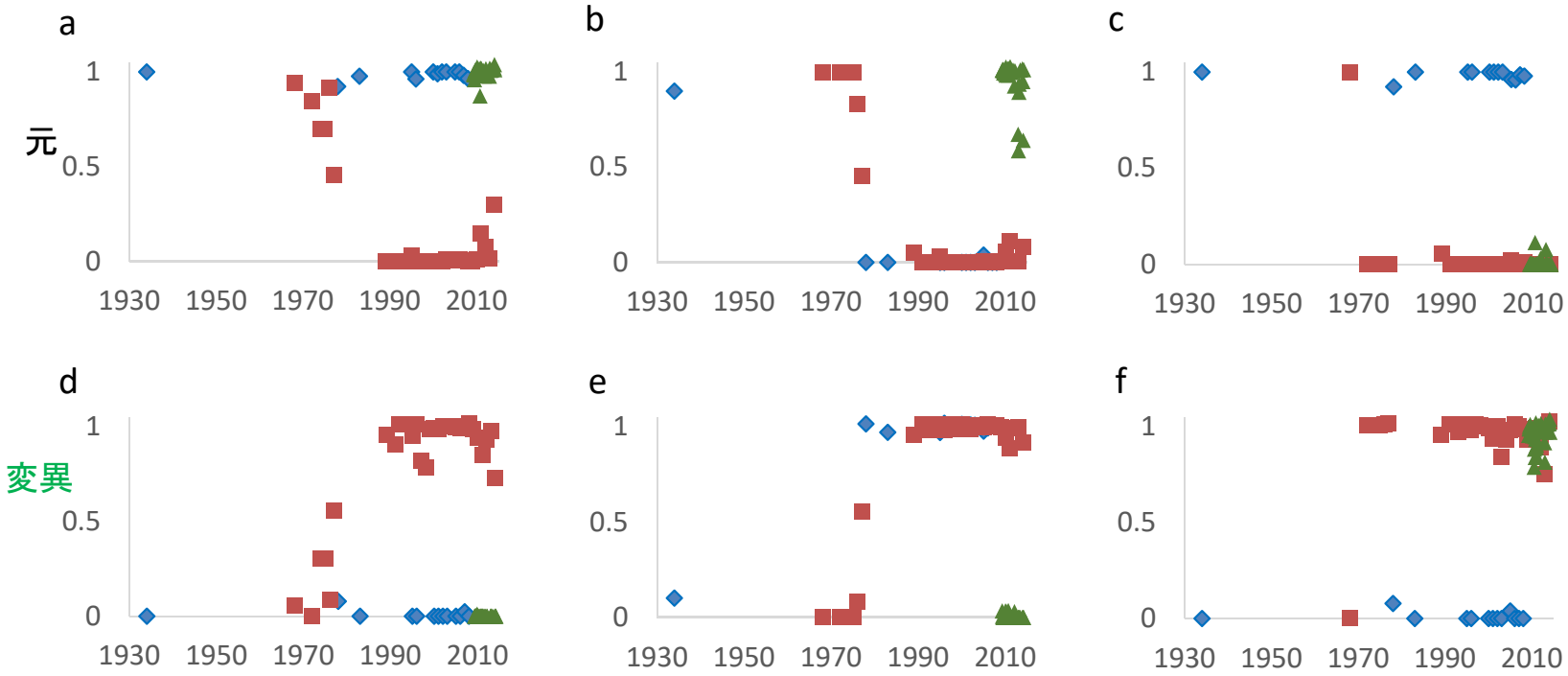
BLSOMによる機械学習でインフルエンザウイルスゲノムの月単位でも観測可能なオリゴヌクレオチド組成の変化が明らかになっていた。

ビッグデータ解析の特徴。まず計算機に聞いてみよう。それから……

Iwasaki et al. *DNA Res* 2011; 18: 125-136.　Iwasaki et al. *BMC Infect Dis* 2013; 13: 386.

待ち構え型の核酸医薬のデザイン。

実験家がデザインしたsiRNA配列の出現率のヒトH1N1とH3N2とpH1N1での時系列変化a-c



混合して使用すると良いと考えられる候補（1塩基変異）の特定. d-f

20-mer（4^{20}) は約1兆1千億の変数

超高次元でスパースなビッグデータからの能率的な知識発見

RNAiについて実験的研究がされていない広範なウイルスを対象にした、AIに支援された解析が重要になる。