

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¹⁾
プロジェクトメンバー 阿部貴志²⁾

¹⁾長浜バイオ大学 バイオサイエンス学部,

²⁾遺伝研・生命情報DDBJセンター

ゲノム塩基配列が未解読な時期にはGC%が、各生物種のゲノムDNAを特徴付ける重要な値であった。しかしこの一変数では広範な生物種の複雑なゲノムを特徴付けるには不十分である。それでは、

2連塩基: AA, AC, -----: 16種類の変数

3連塩基: AAA, AAC, -----: 64種類の変数

4連塩基: AAAA, AAAC, -----: 256種類の変数

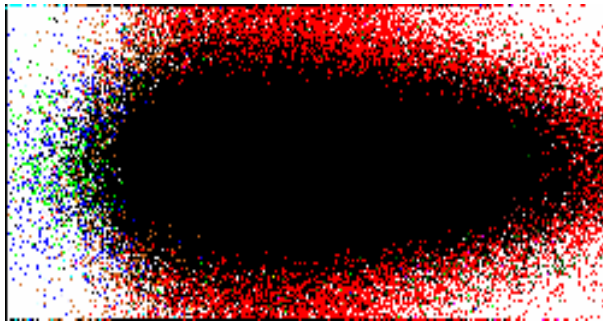
に着目したら何が起きるのか？

このような多くの変数を対象にしながらも、視覚的に理解し易い形式で特徴抽出が可能な情報学的手法のSOM(自己組織化マップ)をゲノム解析に導入。

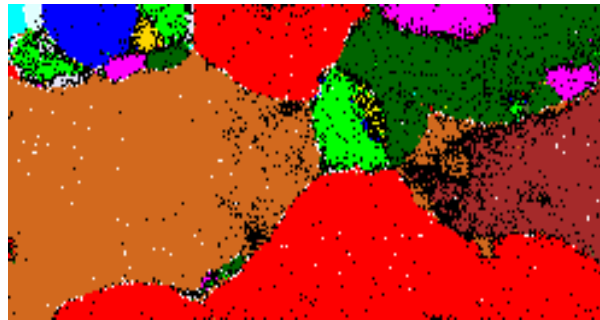
塩基配列が既知な生物のゲノム配列を、10 kbや100kbに断片化してオリゴヌクレオチド頻度の類似度を解析する。

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SOM解析

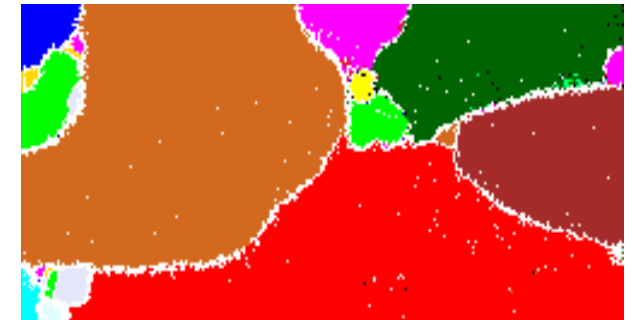
3連塩基PCA, 10-kb



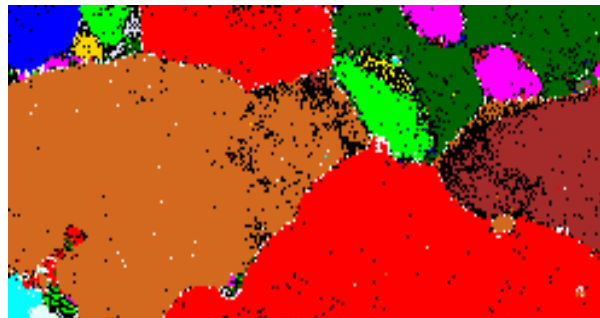
3連塩基SOM, 10-kb



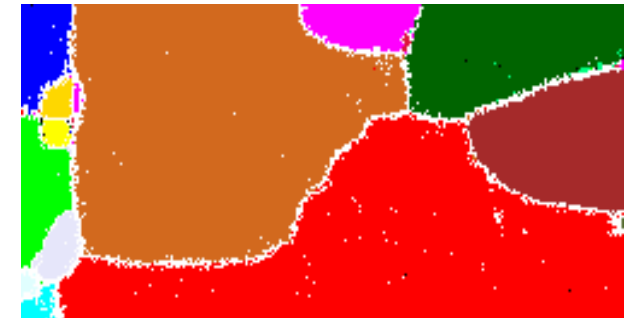
3連塩基SOM, 100-kb



4連塩基SOM, 10-kb



4連塩基SOM, 100-kb



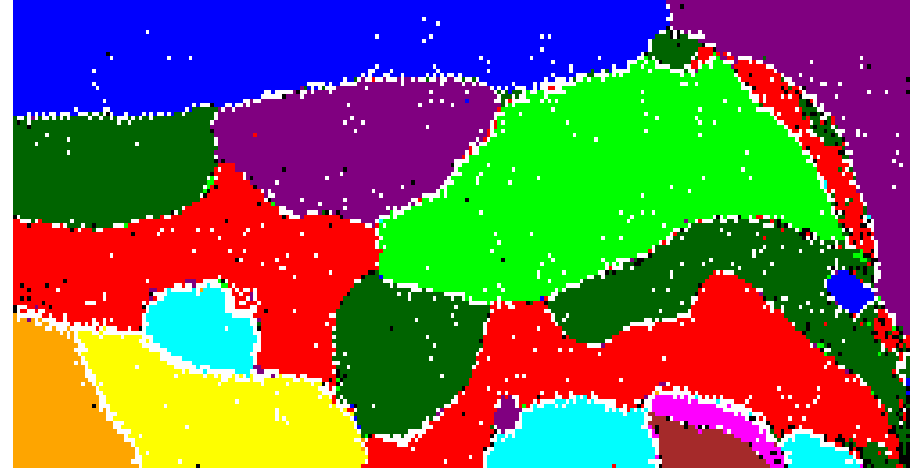
パン酵母 (■), 分裂酵母 (■),
粘菌 (■), 赤痢アメーバ (■),
マラリア原虫 (■), シロイヌナズナ (■),
ウマゴヤシ (■), イネ (■), 線虫 (■),
ショウジョウバエ (■), フグ (■),
ゼブラフィッシュ (■), ヒト (■).

10種類の脊椎動物のゲノム配列を対象にしたSOM

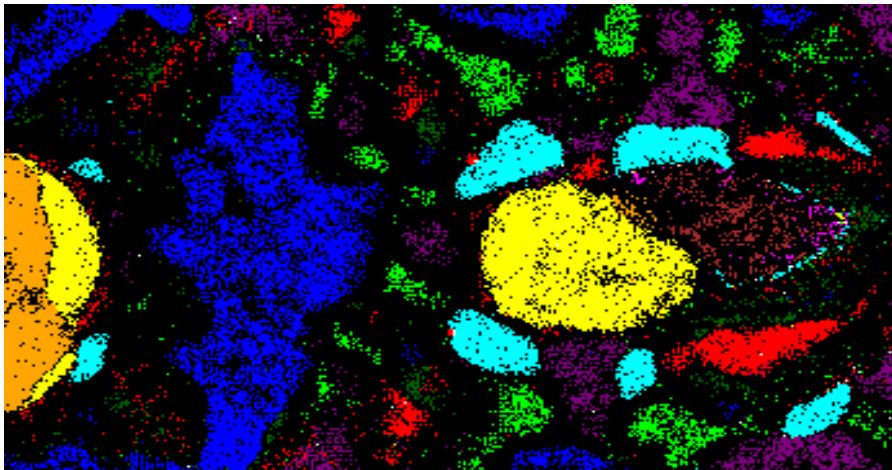
100 kb, 4連塩基SOM



100 kb, 縮退4連塩基SOM



10 kb, 3連塩基SOM



一つ以上の生物種の配列からなる格子点は黒で、一つの生物種の配列のみからなる格子点は以下の色で生物種を示した。

Cow (■), Chicken (■), Dog

(■), Human (■), Frog (■),

Mouse (■), Opossum (■),

Tetraodon (■), Zebra fish (■),

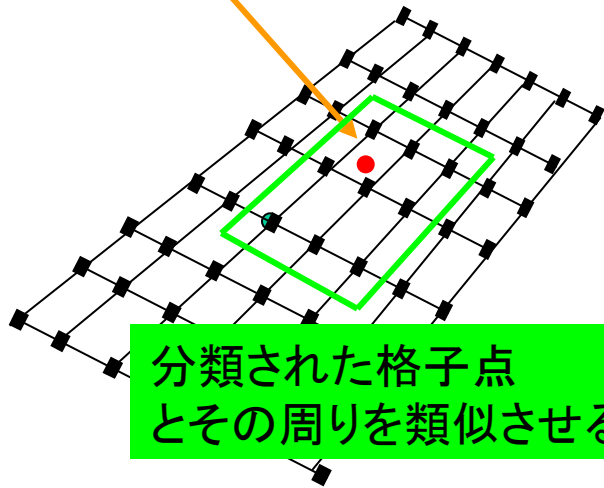
Fugu (■) 縮退では、ゲノム配列の相補性の影響を除去するために、相補的な配列 (例えば、AAAAとTTTT) を同一のものとし、頻度解析を行う。

通常SOM

大統領の原稿10,000件

●, ▲, □, △, ■,

一番近いところを探す



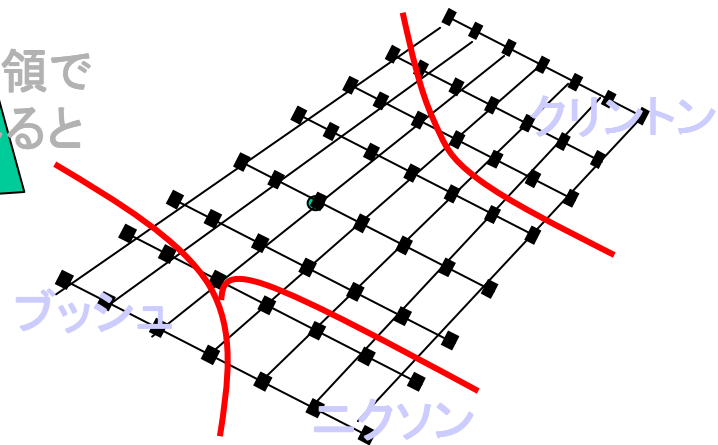
分類された格子点
とその周りを類似させる

例えば20人の大統領の各500件で、合計10,000件の演説原稿について、200種類の単語 (Science, War, Iraq, Peace, Music,) に着目したSOMを考える。

初期状態としては、大統領とは無関係の2,500件の原稿を集めて、各原稿を任意の順番で、50 x 50の2次元の格子点上に並べた上で、200種類の単語の使用頻度を計算する。

$\alpha(t)$ 例えば0.95; $\beta(t)$ 例えば10

学習後、大統領で
結果をみると



50 x 50

2次元上の格子

BL-SOM for genome informatics

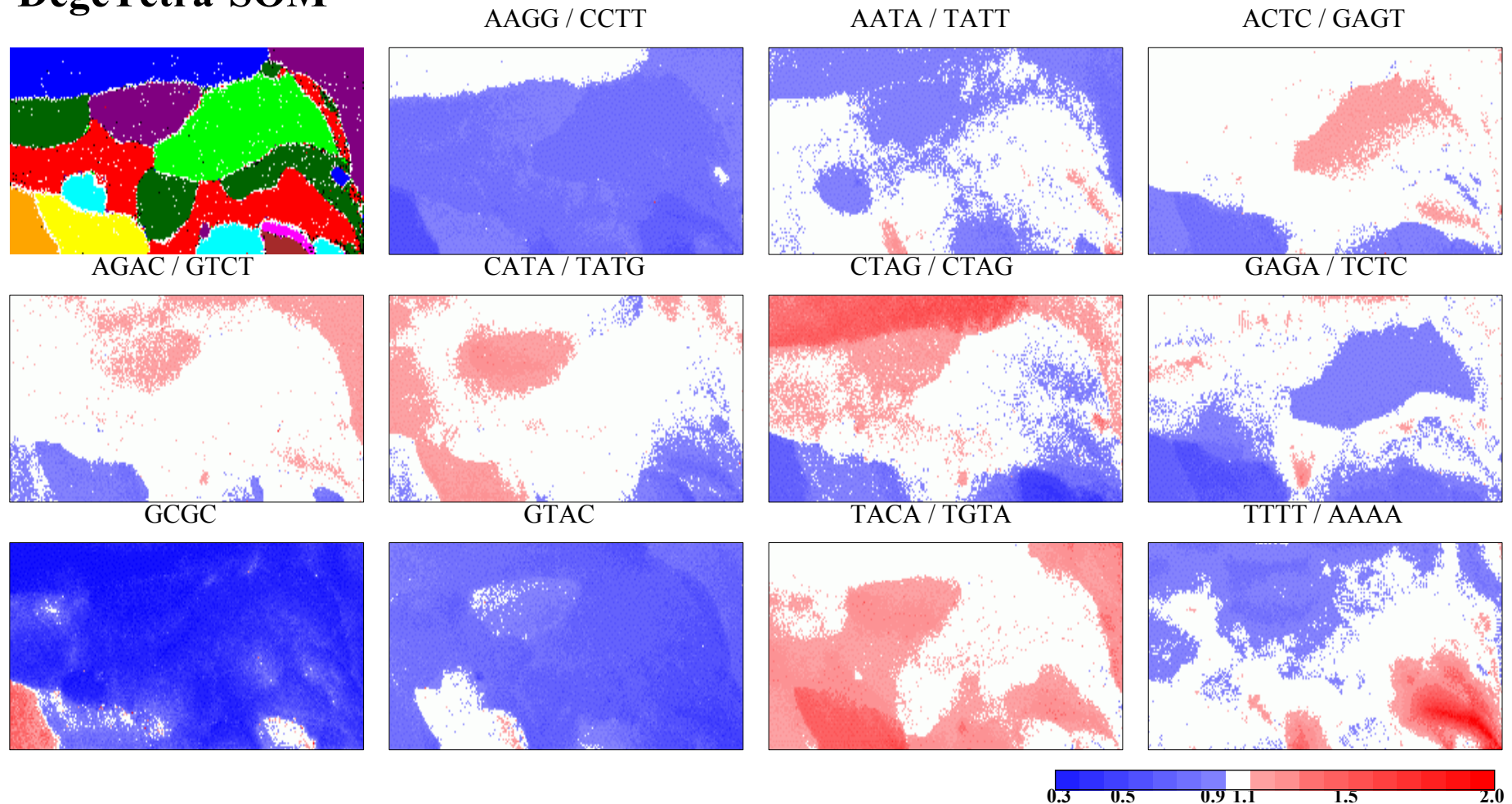
Initial vectors were determined with PCA.

Batch-learning SOM was conducted.

Cow (■), Chicken (■), Dog (■), Human (■),
 Frog (■), Mouse (■), Opossum (■), Tetraodon
 (■), Zebra fish (■), Fugu (■)

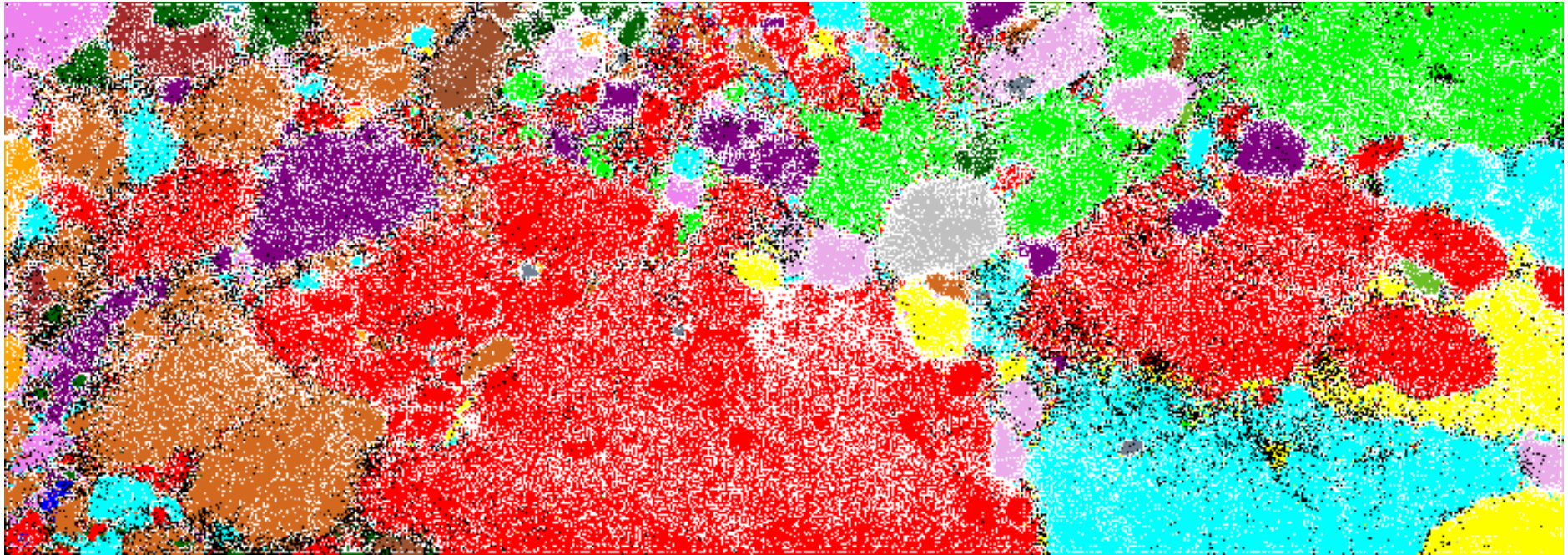
白, ランダム値; ■, 高頻度出現; ■, 低頻度出現

DegeTetra-SOM



膨大な未開拓ゲノム資源 (難培養性微生物ゲノム) の活用

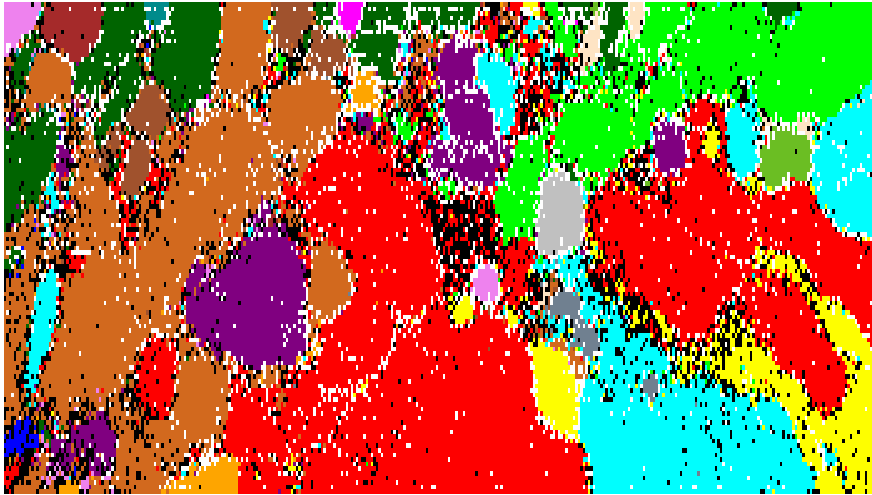
Venterは、Sargasso Sea (near Bermuda) の数百Literの海水由来のDNAのshotgun sequencingを行い、80万本の断片配列(約1 Gb)を得た。120万個の遺伝子の候補を推定している。



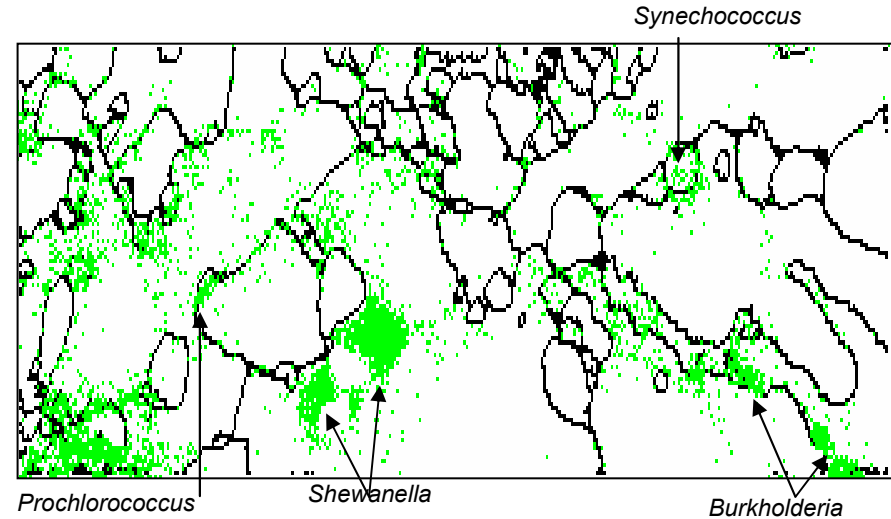
GenBankに収録されている、各ゲノムについて10 kb以上の配列が
解読された全原核生物2,382種(総計約4.5Gb)についての断片化
サイズ5 kb配列の縮退4連続塩基のSOM。縮退では、ゲノム配列
の相補性の影響を除去するために、相補的な配列(例えば、
AAAAとTTTT)を同一のものとし、頻度解析を行う。

Acidobacter (■), Actinobacteria (■), Alphaproteobacteria (■), Aquificae (■), Bacteroidetes
(■), Betaproteobacteria (■), Chlamydiae (■), Chlorobi (■), Chloroflexi (■), Crenarchaeota
(■), Cyanobacteria (■), Deferribacteres (■), Deinococcus-Thermus (■), Deltaproteobacteria
(■), Dictyoglomi (■), Epsilonproteobacteria (■), Euryarchaeota (■), Fibrobacteres (■),
Firmicutes (■), Fusobacteria (■), Gammaproteobacteria (■), Nanoarchaeota (■), Nitrospirae
(■), Planctomycetes (■), Spirochaetales (■), Thermodesulfobacteriales (■), Thermotogales
(■), Verrucomicrobiae (■). 28 Phyla

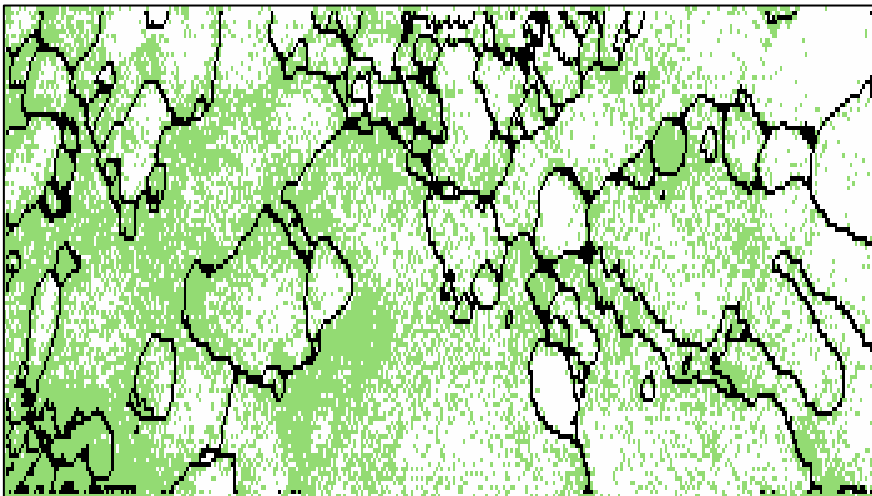
Species-known all Seq.: SOM.



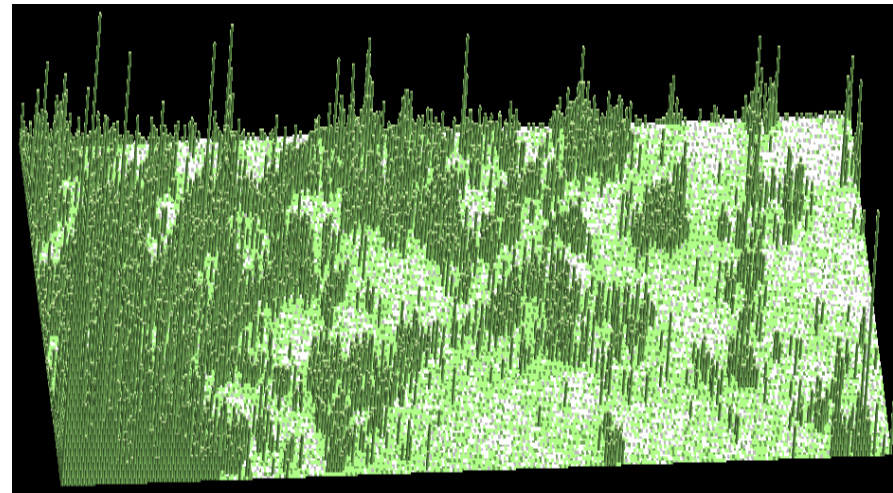
Sargasso Seq. > 5 kb; mapping



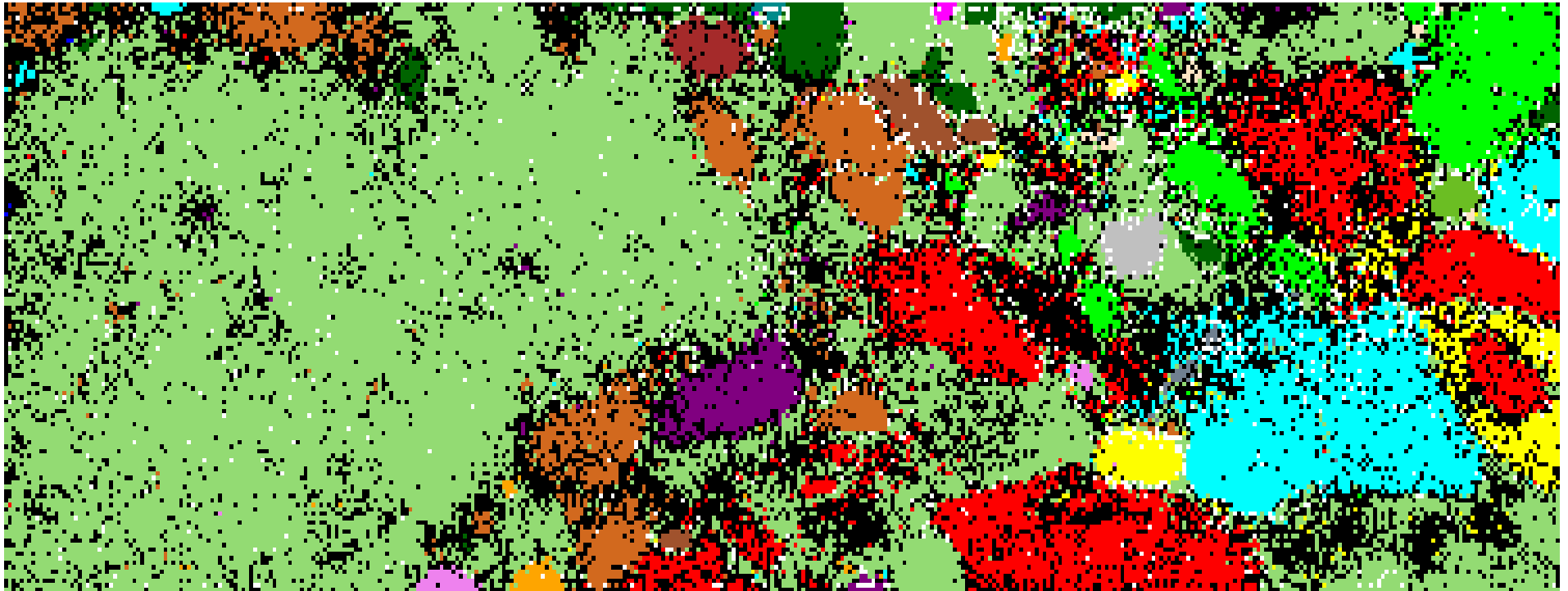
Sargasso Seq. > 1 kb: mapping



All Sargasso Seq.: mapping: 3D



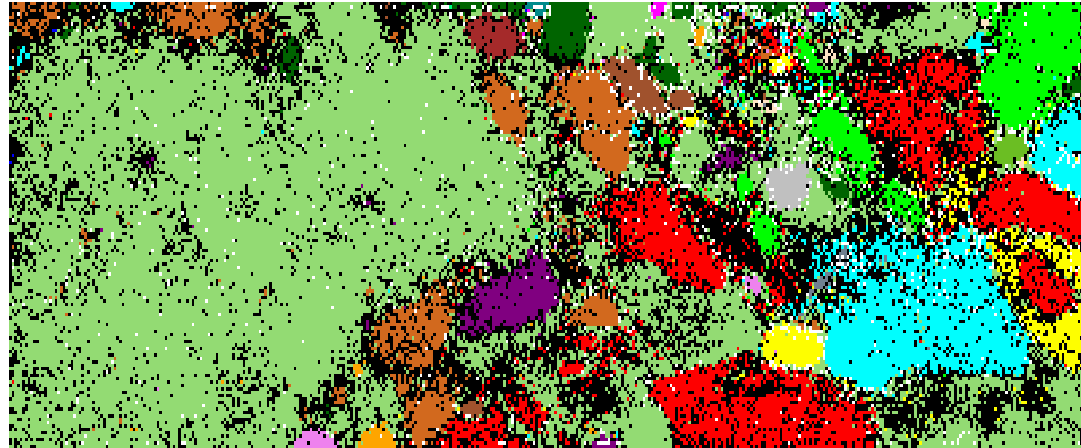
Abe et al., *DNA Res.* 2005



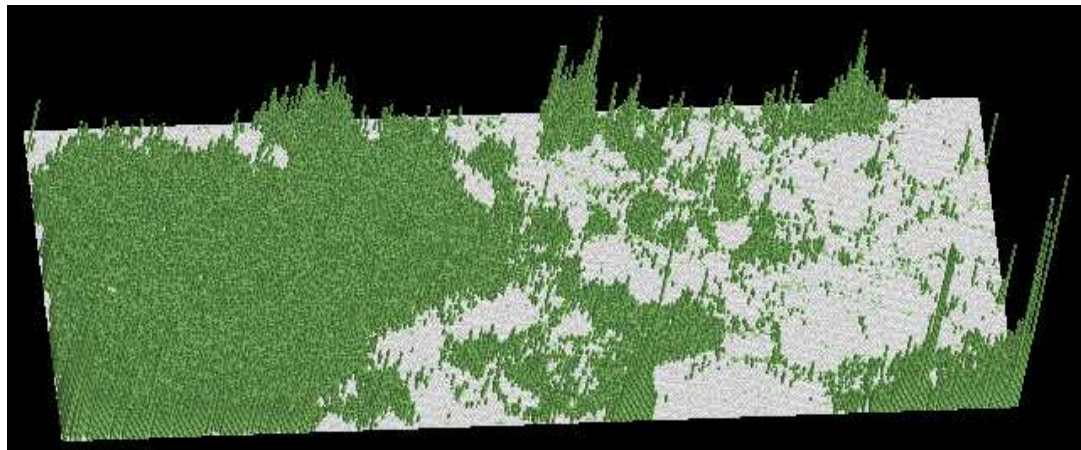
GenBankで10 kb以上の配列が収録されている1,502種 of 原核生物の全配列 (約1 Gb)に、Sargassoの1 kb 以上の長さの配列を加えてSOMを行ったときの結果。

Sargasso配列のみからなる格子点を(■)で示す

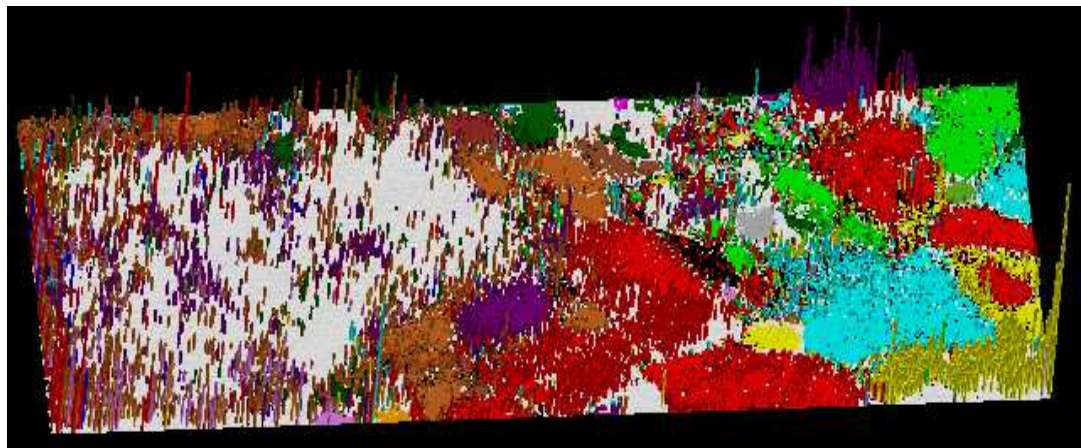
All prokaryote Seq.
plus
Sargasso Seq.



Sargasso Seq.
Unclassified (79%)

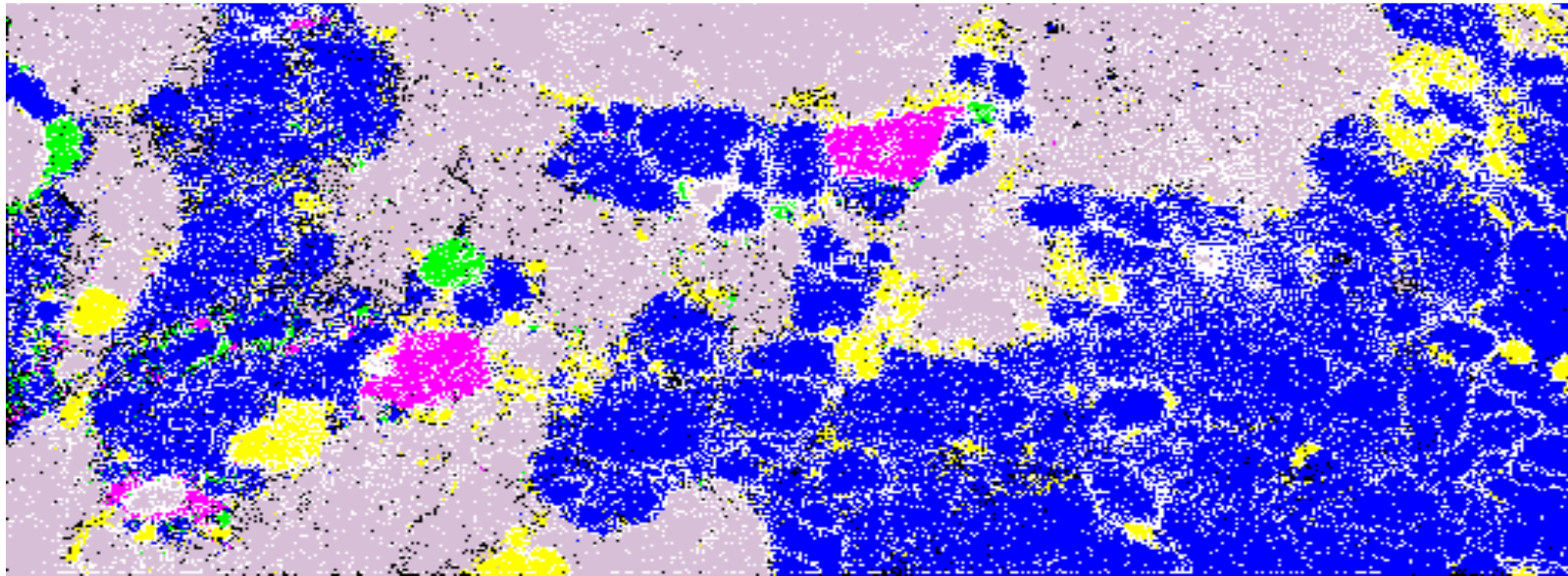


Sargasso Seq.
classified (21%)



Abe et al., *DNA Res.* 2005

Sequences from almost all prokaryotic, eukaryotic, and viral genomes available could be classified according to genomes on a large-scale Self-Organizing Map constructed with the Earth Simulator, *Journal of the Earth Simulator*, **6**, 17-23, 2006.



原核生物1,502種, 真核生物40種, ミトコンドリア 642種、葉緑体 42種、
 ウィルス1,065種での断片化サイズ5 kb, 縮退4連続塩基での大規模
 SOM(地球シミュレータを使用)

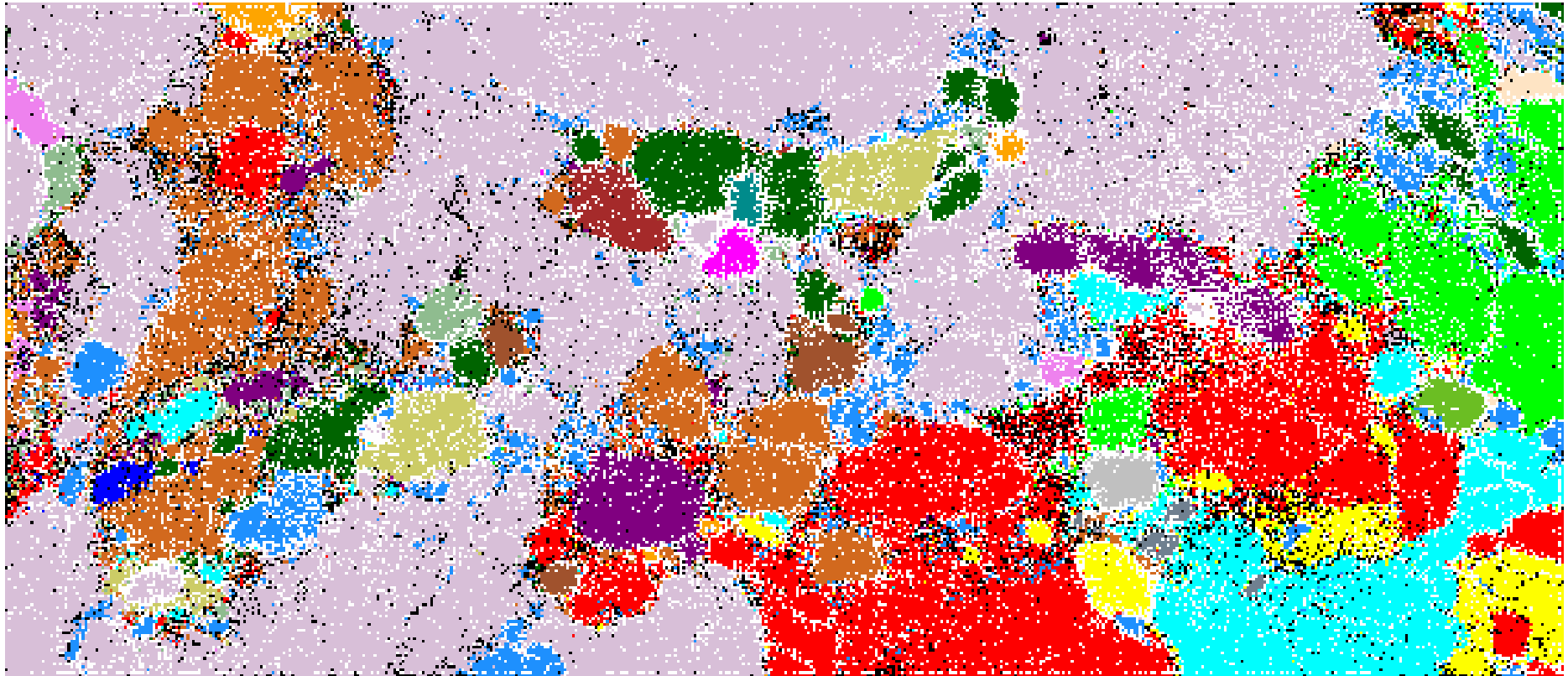
各カテゴリー領域へ分類されていた割合

各カテゴリーの配列数の割合

	Eukaryotes	Mitochondria	Chloroplast	Virus	Prokaryotes	全配列数
Eukaryotes	95.58%	0.19%	0.20%	1.59%	2.44%	180,866
Mitochondria	4.88%	87.04%	3.22%	0.77%	4.10%	8,593
Chloroplast	8.57%	5.24%	72.17%	0.82%	13.19%	6,066
Virus	9.70%	0.21%	0.15%	78.50%	11.44%	32,072
Prokaryotes	1.70%	0.09%	0.35%	1.55%	96.31%	211,567

オルガネラとウィルスの分離能は、80%前後と若干分離能は低い
 が、真核生物と原核生物については、96%と高精度に分離されている

DegTetra: 真核、原核(28 Phyla)、ウイルス、ミトコンドリア、葉緑体



Eukaryote (■), mitochondria (■), chloroplast (■), virus (■), Actinobacteria (■), Alphaproteobacteria (■), Aquificae (■), Bacteroidetes (■), Betaproteobacteria (■), Chlamydiae (■), Chlorobi (■), Chloroflexi (■), Crenarchaeota (■), Cyanobacteria (■), Deinococcus-Thermus (■), Deltaproteobacteria (■), Dictyoglomi (■), Epsilonproteobacteria (■), Euryarchaeota (■), Fibrobacteres (■), Firmicutes (■), Fusobacteria (■), Gammaproteobacteria (■), Nitrospirae (■), Planctomycetes (■), Spirochaetales (■), Thermodesulfobacteriales (■), Thermotogales (■), Verrucomicrobiae (■)

配列相同性検索に依存しないタンパク質の機能推定法の確立を目指して。

配列相同性検索で機能が推定できないタンパク質が100万件を超えるほど大量に蓄積しているが、それらの機能推定を行う方法の開発が急務である。 X線やNMR解析以外の方法

タンパク質の**2連**や**3連アミノ酸**(**400**と**8000**の変数)の使用頻度に関するSOMを行うと、**タンパク質は機能ごとに分離**する傾向を示した。

NCBI-COG(Clusters of Orthologous Groups of proteins)

うち、機能がわかっているCOGのみを使用。

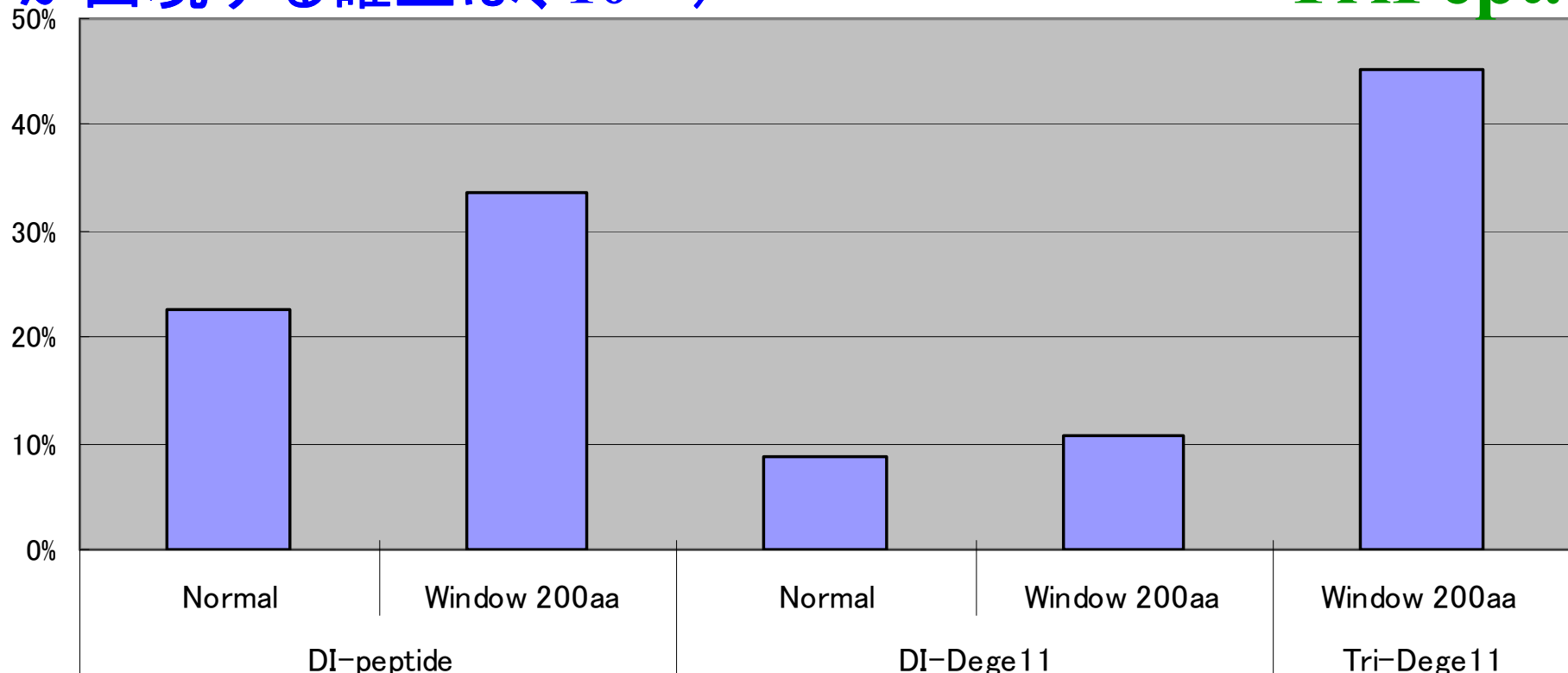
COG数 2,853; 配列数: 111,142

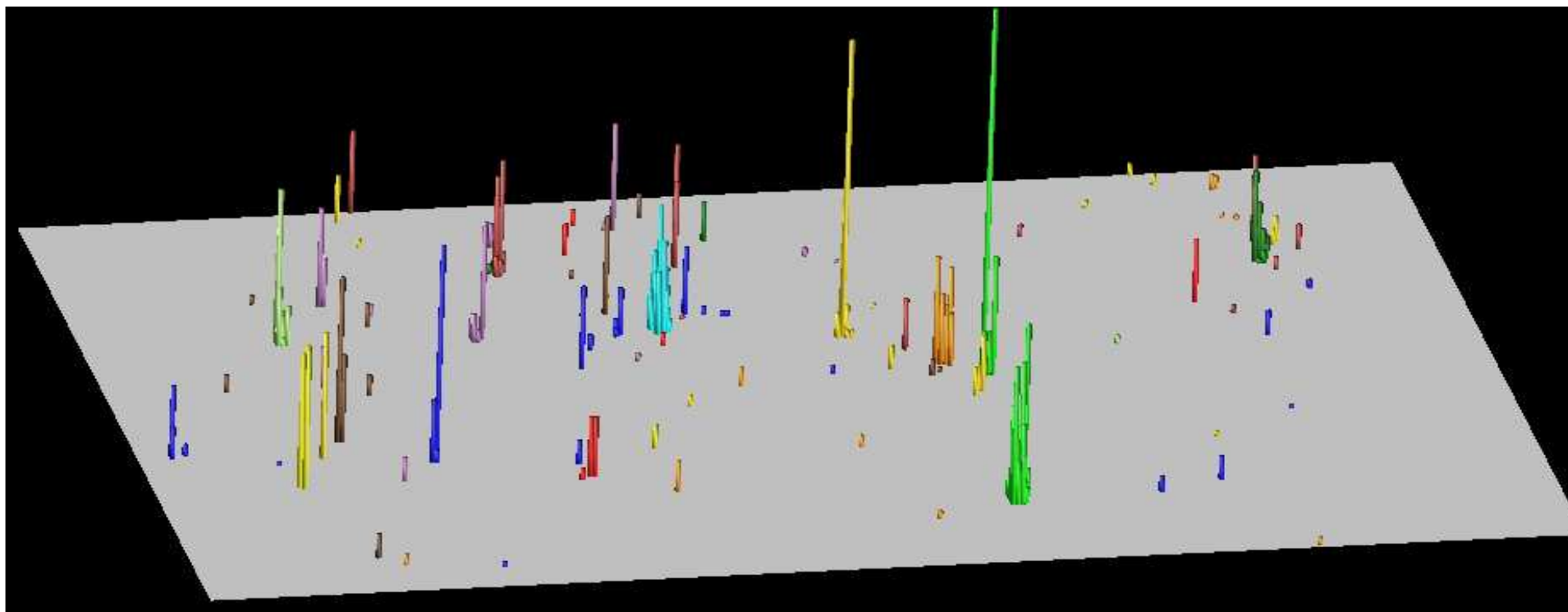
様々な解析条件におけるタンパクSOMの分解能の比較:

COGについて、ピュアな格子点を形成する割合

(格子点あたりの平均の配列数は8; 偶然にピュアな格子点
が出現する確率は、 10^{-28})

TriPept.





機能既知の2853種類の
COGに属する11万件の
タンパク質を200アミノ酸
に分割してSOM解析する
とCOG別に高い頻度で
分離していた。

3連アミノ酸SOM

12個の
COGの例

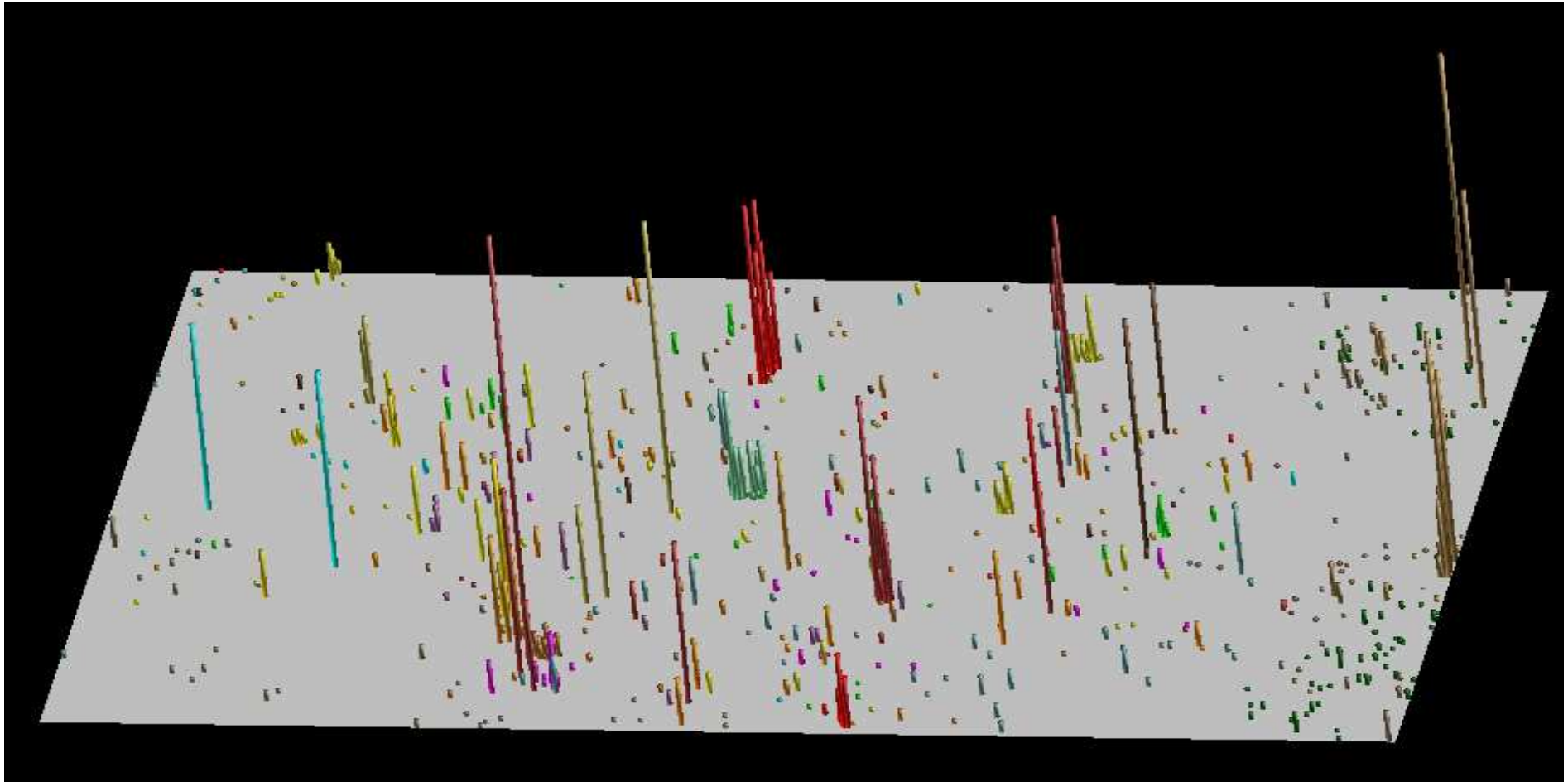
- [J] COG0050 GTPases - translation elongation factors
- [J] COG0090 Ribosomal protein L2
- [E] COG0133 Tryptophan synthase beta chain
- [H] COG0214 Pyridoxine biosynthesis enzyme
- [H] COG0447 Dihydroxynaphthoic acid synthase
- [F] COG0516 IMP dehydrogenase/GMP reductase
- [J] COG0752 Glycyl-tRNA synthetase, alpha subunit
- [E] COG0804 Urea amidohydrolase (urease) alpha subunit
- [M] COG1043 Acyl-[acyl carrier protein]--UDP-N-acetylglucosamine O-acyltransferase
- [D] COG1077 Actin-like ATPase involved in cell morphogenesis
- [M] COG1089 GDP-D-mannose dehydratase
- [C] COG1140 Nitrate reductase beta subunit

1. **NCBI-COG**が混じる格子点を見ると、機能が類似する傾向を示す。

2. **Sargasso-COG**を**NCBI-COG**で作成した3連アミノ酸のSOMへマップすると、その87%が同一のCOGへ帰属していた。

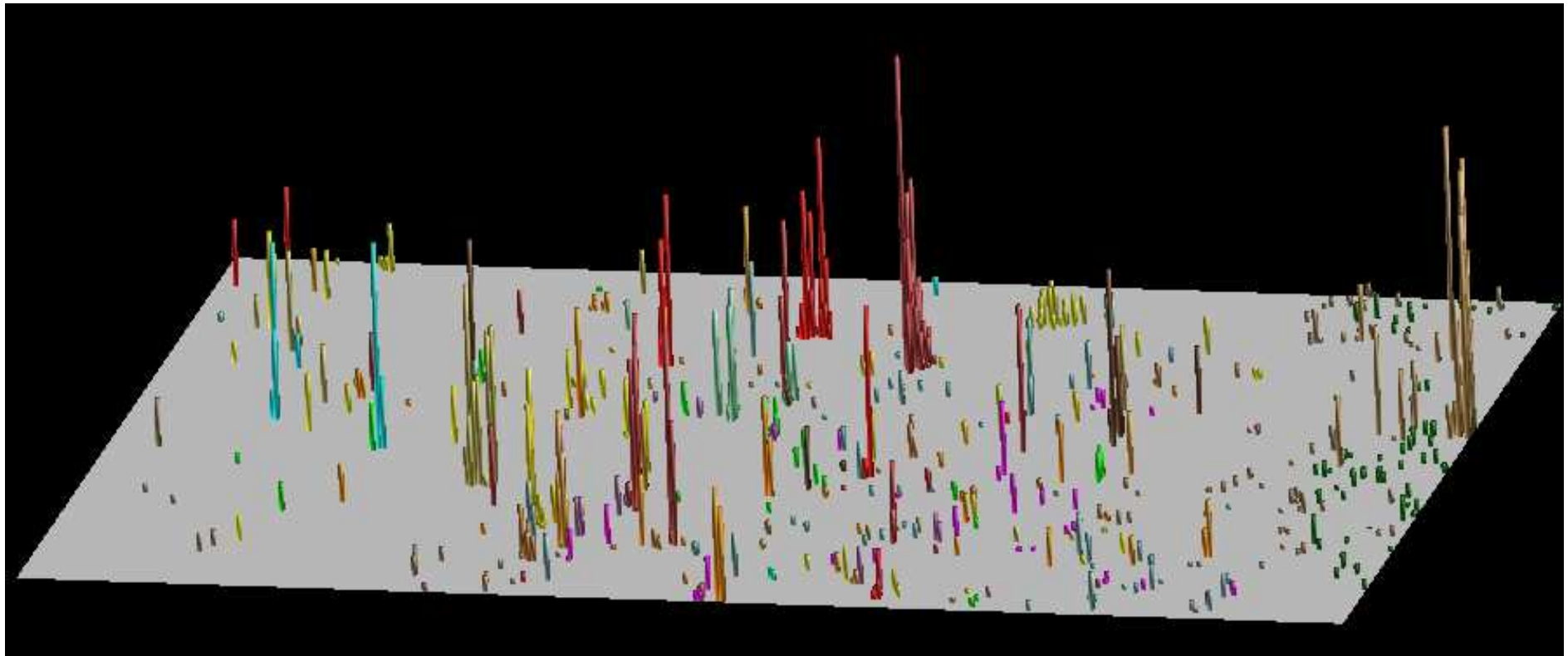
相同性検索でCOGへの分類が不可能な352,066配列の20%近くがSOMによってCOGと関係付けられる可能性が出てきた。2連と3連アミノ酸SOMで同じ答えが出たものから公開の予定。

NCBI-COG 配列でTri11-SOMを作成。 そこへSargasso-COG配列をマッピングした結果



100本以上のSargasso配列が存在するCOGの例について、Sargasso配列の本数を縦棒で表示。21種類のCOGIDごとに別色で示す。

NCBI-COG + Sargasso-COGでTri11-SOMを作成



NCBI-COG と Sargasso-COG が同じ格子点に帰属した場合、Sargasso配列の本数を縦棒で表示。100本以上のSargasso配列が存在するCOGの例。21 種類のCOGIDごとに別色で示す。