自己組織化マップを用いた DNA 配列の解析手法の開発

松藤翔大[†1]　堂薗浩[†2]

概要：近年，多くの生物を対象に実施されているゲノムプロジェクトによって大量の情報が得られている．一方で，得られた大量の情報の中から必要な情報を引き出すことが困難であるという問題がある．本研究では，DNA 配列の塩基の発生頻度を入力データとして自己組織化マップを用いて配列を画像に変換して分類を行うシステムの開発を行った．また，その手法の有用性の比較検討も行った．

キーワード：DNA 配列，自己組織化マップ

Development of analysis method of DNA sequences using Self-Organizing Map

SYOTA MATSUHUJI　HIROSHI DOUZONO

Abstract: Recently, more information was obtained from the genome projects of many organisms. On the other hand, it becomes difficult to obtain meaningful information from the large amount of information. In this study, the method which shows the map of input data comprised of the frequency of occurrence of the bases of nucleotides in DNA sequence with converting the sequences to images was developed using SOM. We also examine the performance of this method in the experiments.

Keywords: DNA sequence, SOM(Self-Organizing Map)

1. はじめに

DNA（デオキシリボ核酸:Deoxyribonucleic Acid）とは，様々な生物の構築と生命活動を維持するのに必要な生物学的情報を含んでいるゲノムを形成しているものである．DNA を解析することで，生物自身の様々な特徴のみでなく，生物で行われている生命活動を調べることができる．また，DNA は，2 つの役割を持っており，1 つ目は，親細胞から子細胞へ遺伝情報を伝えること，2 つ目は，遺伝情報を使って細胞や固体の形や働きを実現することである[1].

自己組織化マップ（Self-Organizing Map：SOM）は，コホネンによって考案されたニューラルネットワークの 1 つで，教師なし学習アルゴリズムの代表的なアルゴリズムである．高次元のデータを低次元データとしてクラスタリングし，可視化を可能とする．SOM の特徴として，入力データ間の類似性が高ければマップ上の近くに配置され，入力データ間の類似性が低ければ遠くに配置される[2].

そこで，本研究では，自己組織化マップを用いて DNA 配列データの解析するためのシステムの開発を行った．DNA 配列は情報学的には大規模な文字列情報であり，そのまま SOM を適用するのは困難である．このような情報を SOM で扱う手法として，情報の統計的情報を用いる方法がある．本研究では統計的情報として文字列の頻度を用いる．このような手法として各 DNA 配列を任意の文字数で区切り単語とし，その単語の発生頻度を DNA 配列から得た情報として，入力データを作成し，SOM によって各配列間の関連性を視覚化して分類を行っていく方法が報告されている[3][4]. 本研究では，この手法とはベクトルの構成を変更し，全ての a文字の組み合わせに対して DNA 配列における発生頻度を入力ベクトルとし，SOM を用いて学習し，各配列の特徴をマップ全体へ写像を行った 2 次元画像として表す方法を提案する．また，この画像をさらに SOM で処理を行うことで，配列間の特徴が分類可能であることを示す.

2. SOM を用いた DNA 配列解析

2.1 DNA とは

DNA（デオキシリボ核酸:Deoxyribonucleic Acid）とは，五炭糖とリン酸，塩基から構成される核酸の一種で，RNA（リボ核酸:Ribonucleic Acid）ウィルスを除くすべての生物に存在している．DNA は 2 本鎖であることが特徴で，1 本の鎖は，4 種類の塩基のアデニン（A），グアニン（G），チミン（T），シトシン（C）が並んで構成されており，また，2 本の鎖の間では，塩基の A と T の間で 2 本の水素結合，G と C の間で 3 本の水素結合を形成して，非常に安定な塩基対を作っている[1].

2.2 自己組織化マップ

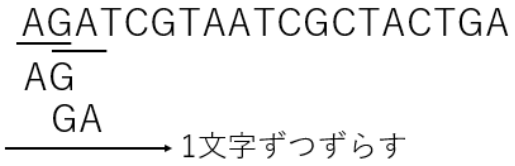
自己組織化マップ（SOM）とは，1982 年にコホネン（Teuvo Kohonen）によって考案されたニューラルネットワークの 1 つで，教師なし学習の代表的なモデルである[2]．SOM は，高次元データの入力情報を二次元のマップ上に写像する．このときデータ同士の距離関係はそれぞれのデータ同士の類似度に依存する．例えば，類似度の高いデータ同士は近い距離に表示され，類似度の低いデータ同士では遠くに表示される．このように，人にはわかりづらい高次元のデータの類似性，関係性をマップ上に写像することで簡単化し，視覚

†1 佐賀大学大学院工学研究科先端融合専攻
†2 佐賀大学

的にわかるようにする．

2.3 特徴量の取得方法

DNA 配列の統計的特徴量として，DNA 配列の塩基の発生頻度を SOM の入力データとして学習を行う方法である[3][4]．実際には，図 1 のように読み取る文字を 1 文字ずつずらしながら数えていく．そして，この得られたそれぞれの塩基の発生回数を，DNA 配列から取得できる発生回数の総数で割った値をその DNA 配列の特徴量として入力データにする．



組み合わせ	AA	AG	AT	AC	GA	GG	GT	GC
数	1	1	2	1	2	0	1	1
入力データ	0.056	0.056	0.111	0.056	0.111	0	0.056	0.056

組み合わせ	TA	TG	TT	TC	CA	CG	CT	CC
数	2	1	0	2	0	2	2	0
入力データ	0.111	0.056	0	0.111	0	0.111	0.111	0

図 1　塩基 2 文字の組み合わせの発生頻度

2.4 実験方法

まず，2.3 に示した方法で DNA 配列から入力データを取得する．各塩基の文字列に対する DNA 配列ごとの入力データを入力ベクトルの各要素として（表 1 の垂直方向のベクトル）SOM で学習する．図 2 のような 1 度目の SOM で得られたマップを各 DNA 配列の入力ベクトルとして SOM で学習することによって DNA 配列の分類を行う．

表 1　2 文字の生み合わせの場合の入力データ

入力ベクトル

	AA	AT	AG	AC	…	GG
イヌ	0.0720	0.0670	0.0540	0.0720	…	0.0770
ホモサピエンス	0.0595	0.0510	0.0560	0.0800	…	0.0825
ネズミ	0.0665	0.0600	0.0490	0.0755	…	0.0740
コメ	0.0630	0.0715	0.0540	0.0685	…	0.0800
ショウジョウバエ	0.0510	0.0705	0.0435	0.0685	…	0.0775
大腸菌	0.0885	0.0600	0.0605	0.0425	…	0.0595

以下に，本手法で DNA 配列を画像へ変換したものを示す．

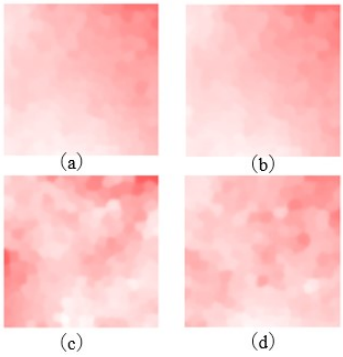


図 2　生物 4 種のマップ

図 2 の(a)はホモサピエンス，(b)はネズミ，(c)は大腸菌，(d)はショウジョウバエを示している．

3. DNA 配列の塩基の発生頻度を用いた解析

3.1 生物種の DNA 配列の解析

この実験では，イヌ，細胞性粘膜，ゼブラフィッシュ，大腸菌，ニワトリ，ホモサピエンス，ネズミ，コメ，ショウジョウバエの 9 種類の生物のアミノ酸代謝系に関連する遺伝子の DNA 配列を 2000 文字ごとに区切って入力データを用意したものを用いて 4 文字の組み合わせの場合と 5 文字の組み合わせの場合を行った．得られた結果を，図 3 と図 4 に示す．

図 3　4 文字の組み合わせ

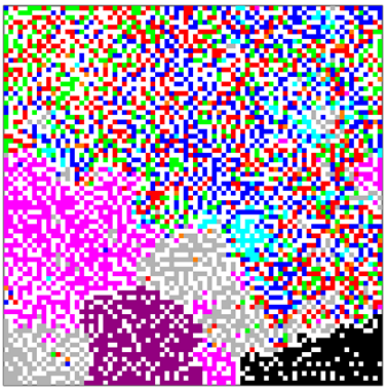


図 4　5 文字の組み合わせ

図 3 と図 4 のマップ上の色は，イヌは緑，細胞性粘膜は黒，ゼブラフィッシュは水色，大腸菌は紫，ニワトリはオレンジ，ホモサピエンスは赤，ネズミは青，コメはグレー，ショウジョウバエはピンクに対応している．

図 3 と図 4 を見てみると，ホモサピエンス，ネズミ，イヌ，ニワトリが交じり合っていて，このことから類似性があることが考えられる．また，そのほかの生物はその生物毎に固まって別れて表示されていることからほかの生物と違う特徴を持っていると考えられる．

3.2 ウィルスの DNA 配列の解析

この実験では，複数のウィルスの全ゲノムデータを用いて 4 文字の組み合わせの場合と 5 文字の組み合わせの場合で行った．用意したデータは，Bacillus，Staphylococcus，Streptococcus，Enterobacteria，Burkholderia ambifaria，Vibrio，Lactococcus，Mycobacterium，Pseudoalteromonas の 9 種類で，得られたそれぞれの結果を図 5 と図 6 に示す．

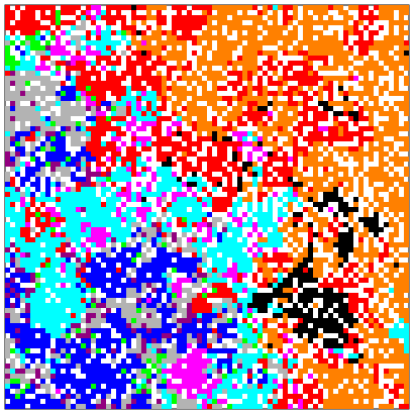


図 5　4 文字の組み合わせ

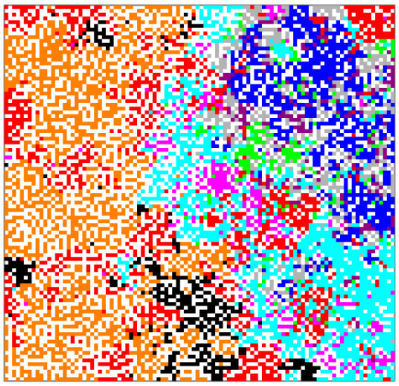


図 6　5 文字の組み合わせ

図 5 と図 6 のマップ上の色は，Bacillus は緑，Staphylococcus は青，Streptococcus はグレー，Enterobacteria は水色，Burkholderia ambifaria は黒，Vibrio はピンク，Lactococcus は紫，Mycobacterium はオレンジ，Pseudoalteromonas は赤に対応している．

図 5 と図 6 を見てみると，各ウィルスの入力データはウィルス毎に 1 つにまとまって別れてはいないが，色ごとに固まって別れているのがみられる．このことから，ウィルスの全ゲノムデータの分類が行われていると考えられる．

3.3 代謝系ごとの DNA 配列の解析

この実験では，ホモサピエンスの代謝系ごとに重複しない遺伝子配列 11 種類のデータを用いて，4 文字の組み合わせの場合と 5 文字の組み合わせの場合で行った．得られたそれぞれの結果を図 7 と図 8 に示す．

図 7　4 文字の組み合わせ

図 8　5 文字の組み合わせ

図 7 と図 8 は，ホモサピエンスの代謝系ごとに重複しない遺伝子配列 11 種類それぞれに色付けして表示している．図 7 と図 8 を見てみると，ホモサピエンスの代謝系ごとのデータから塩基の発生頻度を求めて得た入力データでは分類が行えていないことがわかる．このデータは従来の方法の頻度を用いた方法でも分類が不可能なことが示されており[3]，このことからホモサピエンスの代謝系ごとのデータを分類するにはまた別の手法を考える必要がある．

4. 分類精度の評価

この実験では，本研究の解析手法の有用性を数値的に評価するということを行った．イヌ，細胞性粘膜，ゼブラフィッシュ，大腸菌，ニワトリ，ホモサピエンス，ネズミ，コメ，ショウジョウバエの 9 種類の生物のアミノ酸代謝系に関連

する遺伝子の DNA 配列を用いた．学習データの配列の大きさを 5000，テストデータの配列の大きさを 2000 に区切って入力データをそれぞれ作成する．その後，まず学習データを SOM で学習してマップを作成し，その後，得たマップを用いて，テストデータと比較して，学習データとテストデータの生物種の適合率を数値化することによって評価を行った．その実験によって得たマップを図 9 と図 10 に，適合率の数値を表 2 に示す．

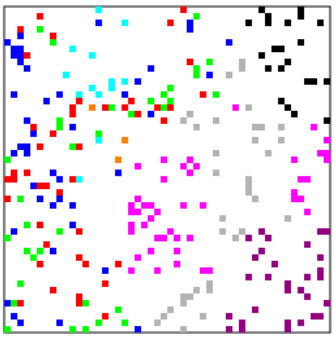


図 9　学習データによるマップ

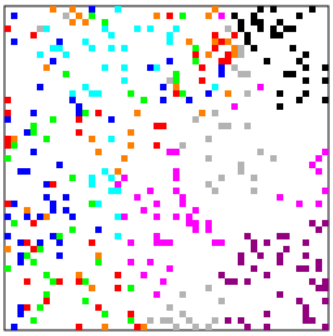


図 10　比較後のマップ

表 2　学習データとテストデータの適合率

生物	[%]
イヌ（緑）	34
細胞性粘膜（黒）	96
ゼブラフィッシュ（水色）	48
大腸菌（紫）	96
ニワトリ（オレンジ）	10
ホモサピエンス（赤）	36
ネズミ（青）	36
コメ（グレー）	74
ショウジョウバエ（ピンク）	80

図 9 と図 10 を視覚的に比較してみると，学習データによるマップとテストデータによるマップの色の配置が似ているのがわかる．このことから，本研究の解析手法の有用性があることを示していると考えられる．また，表 2 の数値による結果を見てみると，細胞性粘膜や大腸菌，コメ，ショウジョウバエといった異なる生物種は値が高いことから有用性を示せていると考えられる．またイヌ，ホモサピエンス，ネズミといった哺乳類 3 種は混ざり合って配置されているのが視覚的にもわかる通り，数値的にも低いものが得られた．また，ニワトリの数値がほかの生物よりも低いのは，この実験で用いているニワトリの DNA 配列がほかの生物の DNA 配列に比べ短いため，十分なデータ個数を取得できなかったためだと考えられる．

5.　結論

本研究では，DNA 配列の塩基の発生頻度を DNA 配列の特徴量として DNA 配列の入力データを取得し SOM による解析手法の開発を行った．各 DNA 配列の塩基の発生頻度を画像に変換し，その画像を SOM で学習することによって DNA 配列の分類を行った．

生物のアミノ酸代謝系に関連する遺伝子の DNA 配列を用いた実験とウィルスの全ゲノムデータを用いた実験では，本研究の手法を用いてある程度の分類が可能であることを示すことができた．しかし，ホモサピエンスの代謝系ごとに重複しない遺伝子配列を用いた実験では，従来の方法でも分類は不可能と示されていたが，本研究の手法でも分類することはできなかった．

分類精度による評価では，似た特長を持つ生物同士では低い数値が得られたが，違う生物種同士では高い値が出ており，生物種ごとの分類はできていると考えられ本研究の手法の有用性が示せていると考えられる．

参考文献

[1] 井出　利憲　著：分子生物のしくみ，
秀和システム，2007
[2] T.コホネン　著，大北　正昭，徳高　平蔵：
自己組織化マップ，シュプリンガーフェアラーク東京
[3] T.Abe, T.Ikemura,et.al, Informatics for unrevealing hidden genome signatures, Genome Res., vol.13m p693-702
[4] 金子　勇太郎：自己組織化マップ用いた
DNA 配列の視覚化と判別の方法，平成 25 年度修士論文