様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　１５　日現在

機関番号：　５５５０１
研究種目：　若手研究(B)
研究期間：　2015～2017
課題番号：　１５Ｋ１６０６６

研究課題名（和文）微生物群集を特徴づける環境要因を推定する自己組織化ニューラルネットワークの開発

研究課題名（英文）Development of a self-organizing neural network for estimating environmental factors characterizing a microbial community.

研究代表者

三澤　秀明（Misawa, Hideaki）

宇部工業高等専門学校・電気工学科・准教授

研究者番号：４０６３６０９９

交付決定額（研究期間全体）：（直接経費）　　2,000,000円

研究成果の概要（和文）：本研究では，自己組織化マップと呼ばれるニューラルネットワークの拡張モデルに基づき，微生物群集と環境の相互作用を明らかにする微生物群集解析手法の開発を目的した．距離型を含む自己組織化マップと高階自己組織化マップにおいて，2つのデータ集合から共通の要因を推定するための新たな学習アルゴリズムを開発した．人工データを用いた実験により，期待通りの結果が得られることを確認した．また，実際の微生物群集データを用いた簡易的な実験において，開発したアルゴリズムの実データへの適用可能性を確認できた．

研究成果の概要（英文）：The objective of this research was to develop a method for analyzing microbial community data based on an extended model of the self-organizing map. We developed a new learning algorithm for self-organizing maps including relational and higher-rank self-organizing maps to estimate common factors from two data sets. The proposed method was applied to artificial data sets and its performance was confirmed. In addition, the proposed method was applied to a real microbial community data set and the possibility of applying the proposed to real data sets was confirmed.

研究分野：　知的情報処理

キーワード：　自己組織化ニューラルネットワーク　異種データ統合　環境要因　微生物群集

1．研究開始当初の背景

近年，生体や自然環境における微生物群集の生態を明らかにすることを目的として，Human Microbiome Project (HMP)や Earth Microbiome Project (EMP)など，遺伝子情報に基づく微生物群集の解析が国内外で盛んに研究されている．試料から抽出した個々の微生物は，個別の DNA 塩基配列で表現され，各地点の微生物群集は DNA 塩基配列の集合として表現される．HMP や EMP などの大型プロジェクトとして，微生物物群集の生態の解明が進められている一方で， そのデータ解析手法については，数えるほどしか研究されておらず，国内ではほとんど見当たらない．

HMP や EMP でも用いられている QIIME や mothur といった解析ツールでは，微生物群集の解析手法として，クラスタリングに基づく手法(OTU 頻度法)や系統樹に基づく手法(UniFrac 法)などが用いられている．しかしながら，これらの方法は，DNA 塩基配列から計算される距離データを，頻度情報や系統樹に一度変換するため，クラスタリング手法や系統樹の作成方法に解析結果が依存するという問題とデータ変換による情報の損失という問題がある．また，群集構造を特徴づける要因の解析には，主成分分析(PCA)や古典的多次元尺度構成法(MDS)などの線形手法が用いられており，非線形性をもつ要因の解析ができていない．

申請者らは，距離データ集合間の類似性を推定・可視化する距離型高階自己組織化マップ(距離型 SOM^2)を腸内細菌群集の解析に適用し，群集構造の類似性を推定できることを示している．距離型 SOM^2 の特長は，情報を歪めることなく DNA 塩基配列から計算される距離情報から群集間の類似性を直接推定できる点である．しかしながら，距離型 SOM^2 による解析では，微生物群集を特徴づける要因を明らかにできていなかった．

2．研究の目的

本研究は，自己組織化マップと呼ばれるニューラルネットワークの拡張モデルに基づき，微生物群集と環境の相互作用を明らかにする微生物群集解析手法の確立を目的とする．具体的には，微生物群集を表す DNA 塩基配列の集合と，多変量の環境データという異なる種類のデータを統合し，微生物群集を特徴づける環境要因を推定する自己組織化ニューラルネットワーク(異種データ統合型 SOM^2) のアルゴリズムを開発する．

3．研究の方法

本研究課題では，異種データ統合型 SOM^2 を開発するために，以下の 3 項目について取り組んだ．

(1) 微生物群集解析手法を評価するためのシミュレーション環境の構築
(2) 微生物データと環境データという異なる種類のデータを統合するための学習方法についての検討
(3) 実際の微生物群集データによる開発アルゴリズムの検証

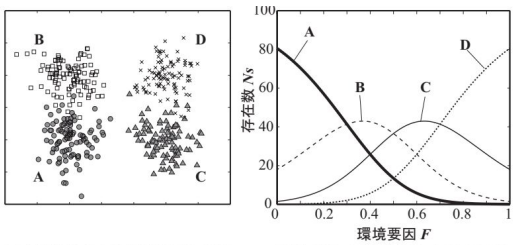
4．研究成果

【平成 27 年度】

平成 27 年度は，(1) 微生物群集解析手法を評価するためのシミュレーション環境の構築と(2) 微生物データ(DNA 塩基配列に基づく距離データ)と環境データ(多変量データ)という異なる種類のデータを統合するための学習方法について検討を行った．

微生物群集解析は，微生物の遺伝子情報と環境データから，環境要因による微生物群集構造(微生物の数と種類)の変化を明らかにする探索的データ解析である．正解が存在しない状況で解析を行うため，解析手法の有効性を定量的に評価できないという問題がある．本研究では，実際の微生物の DNA 塩基配列の距離に基づき，微生物間の距離を定義し，環境要因によって微生物群集構造が変化するシミュレーション環境を構築した．図 1 にシミュレーション方法の基本概念図を示す．

研究申請当初は，このような方法はあまり取られていなかったが，現在ではモックデータセットと呼ばれる人工的なデータを生成する方法として，同様のシミュレーションでの評価が行われている．しかしながら，モックデータセットは，微生物の数と種類を変化させたデータセットを実際の微生物群集データから作り出す方法であり，距離に基づいているデータという点では，本研究の方法はそれらとは異なる．



(a) 微生物間の距離(微生物4種)　(b) 環境要因による存在数の変化

図 1．シミュレーション方法の概念図

学習方法については，距離型 SOM^2 の学習に，環境データを 2 次的な情報として取り入れる方法として，以下の 3 つの方法について検討した．

(a) 距離型 SOM^2 の上位ユニットに 2 次情報を独立して学習させる方法
(b) 距離型 SOM^2 で用いる距離尺度(勝者決定)に 2 次情報を反映させる方法
(c) 距離型 SOM^2 の参照ベクトルの更新式に 2 次情報を反映させる方法

検討の結果，(b),(c)について，引き続き，検討を行うこととした．

【平成 28 年度】

昨年度に引き続き，距離型 SOM² の学習において，微生物群集データに付属する 2 次的な情報を効果的に学習に反映させる方法を検討した．具体的には，2 つのデータ集合から共通する潜在変数を推定する手法である正準相関分析，文書データから潜在トピックを推定する手法であるトピックモデル，データの処理目的に適した距離をデータから学習する距離計量学習などの観点から，異種データを統合するための学習方法について検討を行った．トピックモデルは，潜在的な離散変数を推定する方法として，本研究と関連性がある．しかしながら，潜在的な微生物のグループ（共起する微生物の種類）を見つけることができるが，本研究とは解析の目的が異なった．距離計量（メトリック）学習方法を行う手法は，距離型 SOM² が解析するデータが距離データであるため，メトリック学習を行うことができない点が問題となった．以上の検討の結果，正準相関分析に基づく方法が，本研究の目的に適していることが明らかになった．しかしながら，正準相関分析に基づく方法は，距離データに対しては，直接適用できないため，距離データを取り扱う距離型 SOM² に適した形へ改良する必要があった．

【平成 29 年度】

平成 29 年度は，昨年度までに検討した正準相関分析に基づく方法を，距離型 SOM² に適した形へと改良した．具体的には、距離型を含む自己組織化マップと高階自己組織化マップにおいて，2 つのデータ集合から共通の要因を推定するための新たな学習アルゴリズムを開発した．

提案手法では，同じ対象から得られたデータに対して勝者が一致するように，2 つのデータ集合 X，Y を学習する．図 2 に提案手法の概念図を示す．自己組織化マップまたは高階自己組織化マップが，共有参照ベクトル(近傍係数)を介して，相互に影響を与えながら、交互に学習を行う．

図 3，図 4 に人工データによる開発アルゴリズムの動作確認の結果を示す．共通因子として，1 次元の潜在変数を仮定し，潜在変数から S 状と V 状の 2 次元の人工データを生成した．図 3 は，2 つの SOM を用いて，2 つのデータ集合 X，Y をそれぞれ独立に学習した結果である．独立に学習した場合には，S 状と V 状のデータ分布を近似できておらず，潜在変数を推定できていない．図 4 は，提案手法による学習結果である．提案手法では，S 状と V 状のデータ分布を近似できており，潜在変数を推定できていることがわかる．人工データを用いた実験により，期待通りの結果が得られることを確認した．

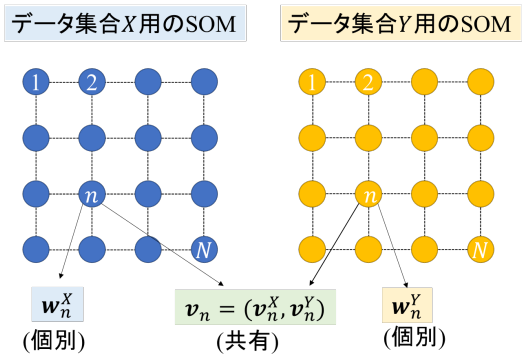


図 2．提案手法の概念図

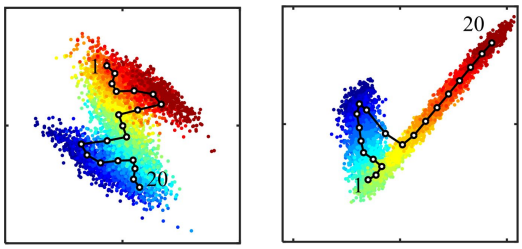


図 3．従来の SOM による学習結果

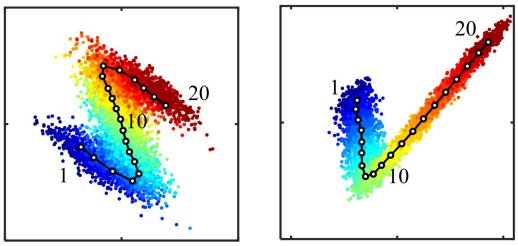


図 4．提案手法による学習結果

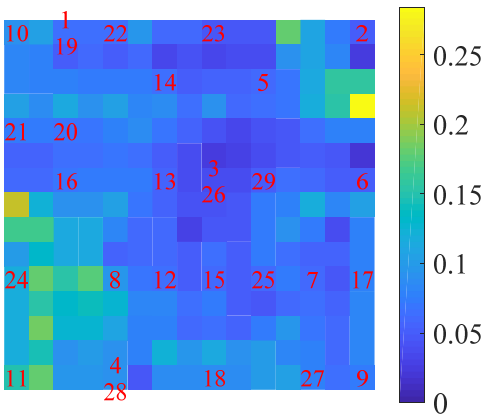


図 5．腸内細菌叢データの解析結果

また，実際の微生物群集データを用いた簡易的な実験において，開発したアルゴリズムの実データへの適用可能性を確認した．微生物群集データとして，健常成人 29 名の腸内細菌叢データと，血液検査データを用いた．血液検査データは，赤血球数，白血球数，ヘモグロビンなどの血液スコア 13 項目と，リンパ球数，CD2 陽性細胞数などの免疫スコア 6 項目から構成される．

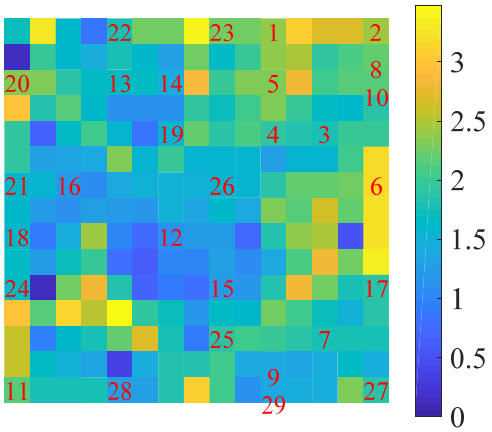


図 6．血液検査データの解析結果

図 5 に開発した学習アルゴリズムによる腸内細菌叢データの解析結果を，図 6 に血液検査データの解析結果を示す．図中の番号は，被験者番号を表す．図 5 は，腸内細菌叢の構造(数と種類)の違いを表しており，図 6 はそれに対応する血液検査データの変化を示している．2 つの結果から，微生物群集の構造と血液検査データの関係性を推測することができる．

今後は，様々な実データを用いた検討が必要である．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計　3　件）

① Hideaki Misawa, Bacterial Flora Analysis by Using Self-organizing Neural Networks, International Conference of Global Network for Innovative Technology (IGNITE) 2016, Penang, Malaysia, January 27-29, 2016.

② Hideaki Misawa, Self-organizing Maps for Extracting Common Latent Variables from Two Datasets, Proceedings of 2018 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing, Honolulu, Hawaii, March 4-7, pp. 239-242, 2018.

③ 三澤　秀明，データ集合間の共通因子を推定する自己組織化マップ，2018 年電子情報通信学会大会講演論文集，A-15-11，東京，2018.

６．研究組織

(1)研究代表者

三澤　秀明（MISAWA, Hideaki）

宇部工業高等専門学校・電気工学科・准教授

研究者番号：40636099