

本件は下記の時間に情報公開が解禁されています。

TV・ラジオ・WEB … 日本時間 平成 30 年 6 月 7 日(木)午前 4 時
新 聞 … 日本時間 平成 30 年 6 月 7 日(木)朝刊

平成 30 年 6 月 19 日

環境と微生物をビッグデータでつなぐ

～「環境」から微生物を検索し、微生物から「環境」を予測するウェブツール「LEA」を開発～

■ 概要

微生物は様々な環境に存在し、物質循環や宿主の健康などに影響しています。したがって「どのような微生物がどのような環境に生息するのか？」という問いは、環境問題や健康問題の解決の糸口となり得るので世界中で探求されました。その結果、特定の「環境」に生息する微生物の群集についてのデータが大量に蓄積される一方で、その「環境」を弁別する記述が研究者間でまちまちとなっていました。この問題によって、研究者間のデータを横断的に解析することが困難となり、微生物の群集と相互作用する環境因子を包括的に理解することができませんでした。

情報・システム研究機構 国立遺伝学研究所の東光一特任研究員および黒川顕教授らのグループは、微生物群集構造⁽¹⁾の大規模データから様々な環境と微生物とのつながりを明らかにし、その結果を利用して環境と微生物のつながりを可視化するウェブツール「LEA」(<http://leamicrobe.jp>)を開発しました。

本研究では、数万サンプルの微生物群集データと各サンプルの由来環境について研究者が記述した文書(自然言語)を確率モデルで表現し、微生物群集の形成に影響を及ぼす環境を機械学習技術によって抽出しました。抽出された「環境」はそれぞれ固有の「微生物」群集構造と結びついています。この結果をもとに、「環境」に対応する「微生物」を検索する、あるいは逆に「微生物」から「環境」を予測するウェブツール「LEA」を開発しました。LEA は、ヒト共生微生物による疾患の診断や、環境汚染の検出などへの応用が期待できます。

本研究成果は、「PLOS Computational Biology」に平成 30 年 6 月 6 日午後 2 時(米国東部標準時間)に掲載されました。

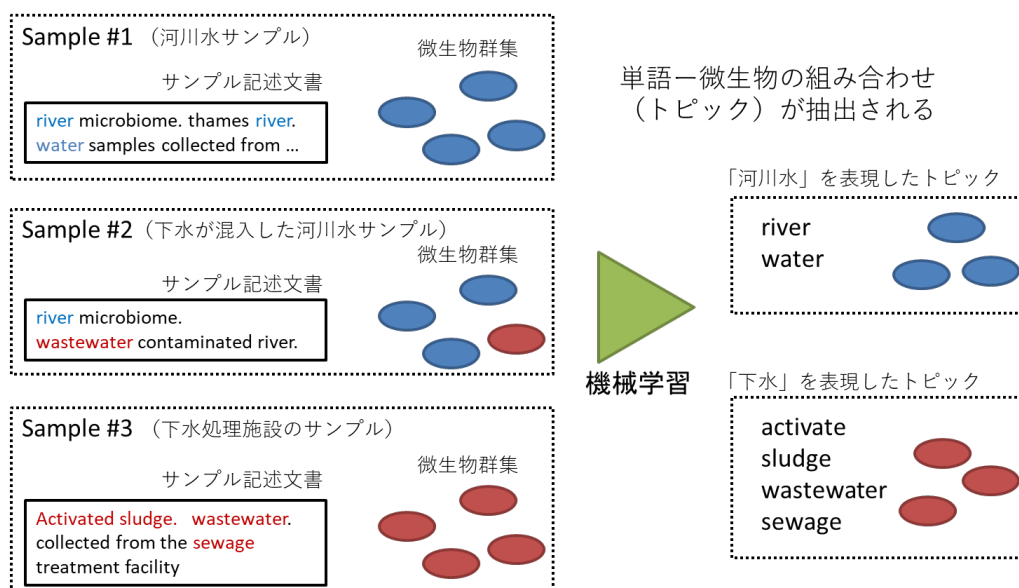


図 1: 微生物と自然言語の対応関係を学習

DDBJ などの公共データベース中のすべてのメタゲノムデータには、「サンプルについて研究者が記述した文書(自然言語)」が付随する。

■ 成果掲載誌

本研究成果は、米国電子ジャーナル PLOS Computational Biology に平成 30 年 6 月 6 日(米国東部標準時間)に掲載されました。

論文タイトル: Latent environment allocation of microbial community data (微生物群集構造データにおける潜在的環境因子の推定)

著者: Koichi Higashi, Shinya Suzuki, Shin Kurosawa, Hiroshi Mori and Ken Kurokawa (東光一、鈴木真也、黒澤晋、森宙史、黒川顕)

■ 研究の詳細

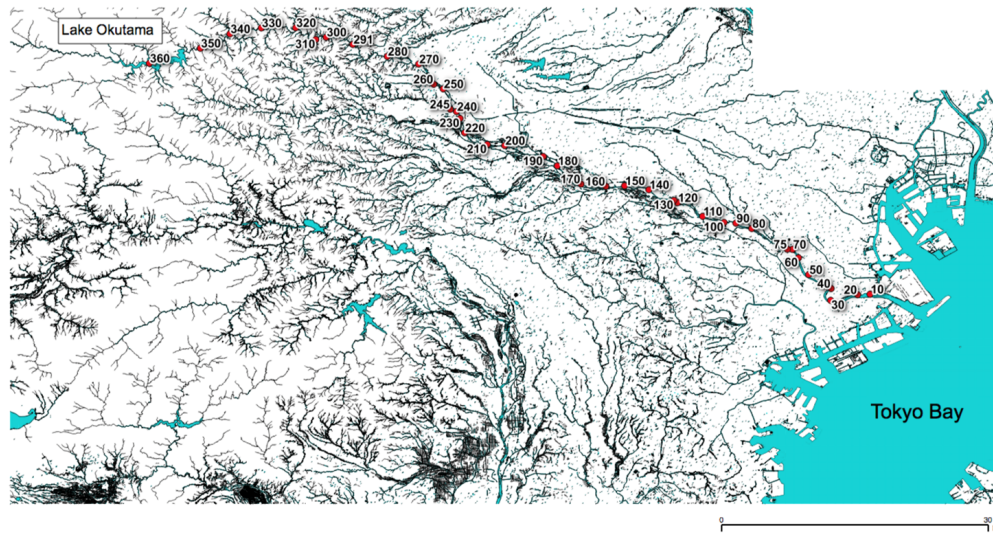
● 研究の背景

近年、世界中の様々な環境に生息する微生物の群集構造データが爆発的な勢いで蓄積されています。微生物群集はその生息環境と密接に結びついているため、微生物群集の変化の予測や制御が可能となれば多くの産業に有用な技術となりますが、微生物群集形成に影響を及ぼす環境因子はあまりよくわかっていません。

これまでの研究では、まずサンプルを環境ごとに分類し、環境間の微生物群集の差異を調べることで特定の環境に関連する微生物を明らかにしてきました。しかしながら、ヒト腸内環境など、個人によって異なる微生物群集構造をもたらす些細な環境の違いもあれば、河口のように、河川と海洋との境界の定義が難しい環境もあります。さらに、そもそも環境に関する記述がないサンプルも多く存在します。したがって従来のように個別に環境カテゴリを各サンプルに割り当てる手法では、きわめて多様な環境に由来する大量のサンプルを包括的に解析することは困難でした。

● 本研究の成果

本研究では、微生物群集構造と、その由来環境に関する文書(自然言語記述文書)が対となった 3 万以上のサンプルを含むデータセットに統計的潜在意味解析(トピックモデル⁽²⁾)を適用し、どのような「環境」が存在するのかをデータから自動的に抽出しました。抽出された「環境」(トピック)は、共起微生物と共起単語セットとの対応関係を表し、群集構造と文書の両方に影響を与える潜在的な因子と考えられます。サンプルは複数トピックの混合として表現されるため、全サンプルのトピックに基づく類似性の可視化により、多様な環境の連続性を評価できます。これらの機械学習結果を利用して、膨大な数のサンプルを探索するウェブツール LEA(Latent Environment Allocation)を開発しました。新たに取得したサンプルの「微生物群集構造データ」をアップロードすることで、LEA はサンプルの「由来環境」を高速に予測し、その微生物群集構造の典型性・異質性を評価することも可能です。



多摩川38地点細菌群集の系統組成
河川上流

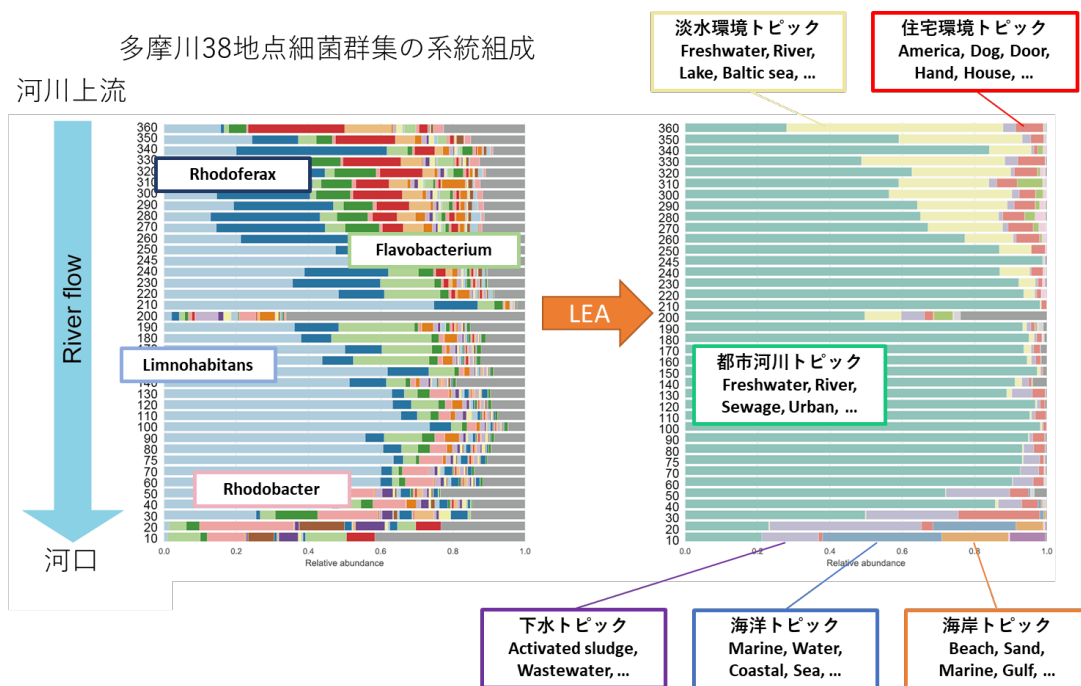


図 2: 微生物群集構造データからの環境予測

テスト対象: 多摩川細菌群集

上流(奥多摩湖)から河口(羽田空港付近)まで 38 サンプルを解析

● 今後の期待

LEA は、新規サンプルにおける微生物群集構造の性質を高速に評価し、由来環境において典型的に見られる微生物群集構造からの「逸脱」を的確に検出するシステムです。したがって、ヒト共生微生物を利用した疾患の新たな診断や、環境汚染の検出など、微生物群集の分析、制御などを目的とした事業全般への応用が期待できます。

■ 用語解説

(1) 微生物群集構造:

微生物は地球上のほとんどあらゆる環境に存在し、多種多様な微生物が群集を形成して生息している。微生物群集はその内部の複雑な相互作用を介して全体としてその生息環境に大きな影響を及ぼすため、どのような系統の微生物がどれほどの割合で存在しているか(群集構造)を知ることが重要である。過去 10 年、DNA

シーケンシング技術の発展とともに、培養に依存しない群集構造解析手法（系統マーカー遺伝子を対象とした群集構造解析やメタゲノム解析）が広く利用され、様々な環境中の微生物群集構造についての理解が飛躍的に進展した。

(2) トピックモデル：

主に自然言語処理の分野で大量の文書データを対象としてテキスト内の潜在的な（直接観測することのできない）意味構造を発見するために利用されている確率モデル。それぞれの文書はなんらかの「トピック」がある割合で混合されて生成されたものであると仮定し、トピックは文書データの集合から単語の出現頻度を元にして推論される。トピックは意味的に類似した単語のクラスターである。たとえば「バイオインフォマティクス」について記述した文書は、「生物学トピック」に関連した単語と「情報科学トピック」に関連した単語が混ざって構成されていると考えられる。

■ 研究体制と支援

本研究は、文部科学省科研費「新学術領域研究『学術研究支援基盤形成』」先進ゲノム解析研究推進プラットフォームの助成を受けて行われました。

本研究成果は、国立研究開発法人科学技術振興機構バイオサイエンスデータベースセンター（JST-NBDC）のライフサイエンスデータベース統合推進事業（統合化推進プログラム）における研究開発課題「データサイエンスを加速させる微生物統合データベースの高度実用化開発」の一環として開発されました。

■ 問い合わせ先

<研究に関すること>

- 情報・システム研究機構国立遺伝学研ゲノム進化研究室
教授 黒川 顕（くろかわ けん）

<報道担当>

- 情報・システム研究機構 国立遺伝学研究所 リサーチ・アドミニストレーター室
清野 浩明（せいの ひろあき）