

[研究助成 (A)]

匂い情報空間の構築, 分子科学と機械学習の融合による挑戦

Building molecular odor sensing space, a study for the fusion of scientific computing and machine learning

2221022



研究代表者

熊本大学 大学院先端機構

特任助教

原 田 祐 希

[研究の目的]

嗅覚機能が低下すると食品の腐敗や火災の発見が遅れ, 日常生活の危険が増え, 精神的な健康も阻害される。本研究では, 匂いの情報空間の基盤構築を進める。匂いの情報化が波及する分野を図1に示す。健康や医学のための科学にとって匂いは重要であるだけでなく, 美味しさは「フレーバー」で感じられ, また美しさを「フレグランス」として想像できる。つまり人間らしい文化にも匂いは重要である。

匂いは, 各種センサ, 分析機器 (GC, LC (GCMS 含む), MS), 種々の動物における嗅覚受容体の包接挙動, 嗅覚受容細胞群の神経応答, 嗅球の活性化, 脳の刺激による活性化, ヒトによる官能評価によって, 数値化されたり, さまざまな言語的な表現になる。本研究者がこれまでに提案した「匂いセンシング空間」は, センサおよび分析機器による手法, 生理学的・分子生物学的手法, あるいは官能評価によって匂いが受ける多次元情報化において, 得られる情

報空間である。匂いを情報化するためには“感覚量”に近い匂いセンシング空間の設定と解析方法の確立が必要である。

匂いの質の記述子には冗長性があり, 匂いは遺伝的および文化的背景や個人差経験を反映し知覚される。匂いの情報化の最初段は, 約 400 種類の嗅覚受容体と約 40 万種類の匂い分子との分子間相互作用であるが, 分子間相互作用の観測やシミュレーションはまだ困難だ。

本研究は匂い物質を解析対象とし, 匂い情報空間の構築を目的としている。これは匂いの質を解析する前提をもたらし, 匂いセンシング・情報化・VR 再現におけるインターフェース設計への貢献が期待できる。

[研究の内容, 成果]

フィンガープリントによる匂い情報空間の構築

本検討では「匂い分子の情報空間」と「匂い記述の情報空間」について回帰分析を行い, 各匂い記述子に該当する化合物の集団 (ここでは各“サブ”ケミカルスペースと呼ぶ) 個々の「匂い記述」の予測について検証・比較した。

検討には, 700 種類の単品ケミカルについて 196 の評価用語による官能評価が収載されている Flavornet データベースを使用した (<http://www.flavornet.org>)。官能評価は 196 の用語 True/False によって評価され, 複数の評価用語が該当する化合物もある。回帰分析を十分で

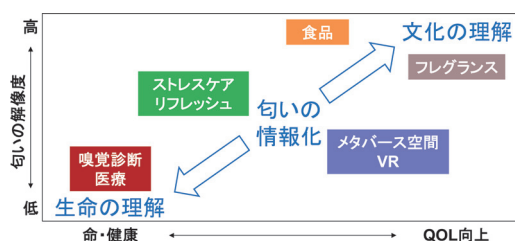


図1 匂いの情報化が波及する分野

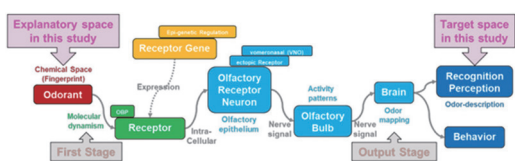


図2 本研究の着目する匂いセンシング空間

きよう該当化合物が8分子以上の匂い記述を解析対象とし、匂い分子の分子構造から「匂い記述」の予測について検証した。

分子の記述子は1世紀以上にわたって様々に提案されてきたが、匂いの官能評価を十分に予測することはできないとされてきた。図2に示すように本検討の回帰の説明変数は、計算科学ソフトウェア MOE およびオープンソースソフトウェア RDkit によって得られる広く用いられているフィンガープリントであり、分子構造を表すベクトル表現である。匂い分子の分子構造は CAS ナンバーにより取得した。回帰の目的変数は匂い記述である。この結果、予測能力はサブケミカルスペースごとに異なることがわかった。

匂い記述子に該当する各サブケミカルスペースの比較にも、計算科学ソフトウェア MOE およびオープンソースソフトウェア RDkit によるベクトル表現を使用した。まずタニモト係数を用いて集団内の類似度の中心値・分布の確認し、ついでフィンガープリントによる主成分マッピングすることでサブケミカルスペースを比較した。

図3, 4はこの比較の結果の一部である。各(A)に検討を行った予測のROCを、AUCを青文字で示す。各(B)に集団内全ペアのタニモト係数のヒストグラム、各(C)が全化合物のフィンガープリントの第1第2主成分によるマッピング(曲線は連続的な存在確率を仮定したカーネル密度推定を表す)を示す。両図に示す検討はフィンガープリントに ECFP (Extended Connectivity Fingerprint) を用いた。

評価用語“must”, “oil”, “fresh”の指すサ

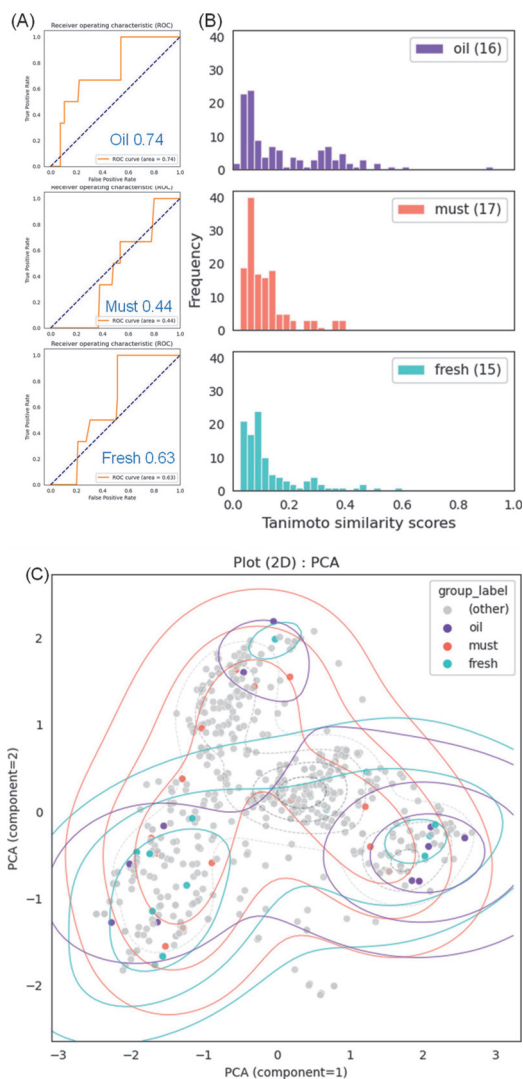


図3 “AUC 0.7 以下 (予測不可)” の記述子に対応するサブケミカルスペース

ブケミカルスペースは集団内類似度が低く、それぞれが似ていないサブケミカルスペースの例である。図3に示すように、集団内タニモト係数(B)は小さめであり、マッピング(C)におけるプロットは離散的である。これらのAUC(A)はおおよそ0.7以下と、予測精度は低い。

評価用語“rose”, “cabbage”, “cucumber”の指すサブケミカルスペースは集団内類似度が高く、それぞれが似ているサブケミカルスペースの例である。図4に示すように、集団内タニモト係数(B)は大きめであり、マッピング(C)におけるプロットは集散的である。これらのAUC(A)はおおよそ0.9以上と、予測精度は高い。

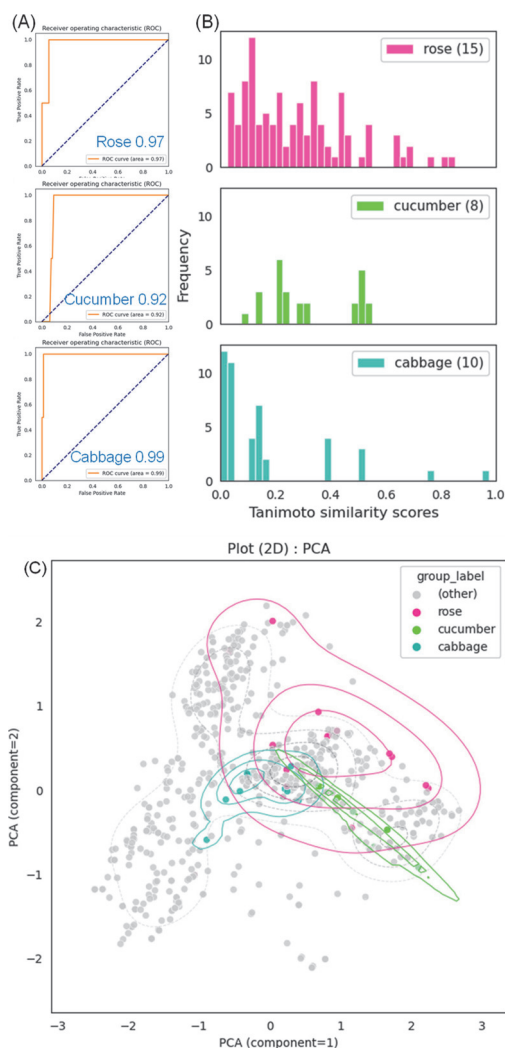


図4 “AUC 0.9 以上（予測可能）”の記述子に対応するサブケミカルスペース

以上のように評価用語の指す化合物群が似ていると予測精度が高いことがわかった。さらに汎用に使用される評価用語は、マッピングにおけるプロットは離散的で予測精度が低めである。これはサブケミカルスペース内の分子構造が多様であることを示唆する。同検討を他のフィンガープリントでも検証したところ、その差異よりも、サブケミカルスペース内の分子構造の多様性が大きく予測性能に影響した。

（本結果は投稿中である）

分子科学計算にもとづく匂い情報空間の構築

分子記述子は1世紀以上様々に提案されてきたが、分子構造の情報空間を十分網羅できてい

ない。本研究では匂い情報空間について、(I)既存の化学情報学に基づく記述子、(II)新しい機械学習を適用した分子記述子、(III)分子科学計算に基づく記述子、この3者による情報空間の構築に取り組んだ。

受容体ないしリガンドに関する分子科学計算（分子軌道計算、分子動力学）は古くからおこなわれており、生体分子の系は平衡ではなく、分子動力学計算や時間発展の分子軌道計算といった非平衡の計算手法が必要とされる。嗅覚受容体と匂い分子の相互作用はポテンシャルと自由エネルギー差(ΔG)により解釈される。これらは現在も議論が行われていると同時に、計算機実験として分子科学において重要な手法となっている。種々提案されている高精度計算を最適な計算条件で実行し、最新の分子科学計算に基づき解釈することが必要である。

本検討ではマルチ計算タスク、大量のデータ処理が可能な高性能サーバーを用意し、MOEおよびGaussianを用いてFlavorNetデータベースに記載されている700種類の匂い分子について分子振動計算に注目した量子化学計算を実行し、解析を行った。（本検討は継続して検討を行っている）

グラフの機械学習を応用した計算科学にもとづく匂い情報空間の構築

深層学習では、これまで個々のデータの集まり（ベクトル）や、格子状に配列されたデータ（画像）、逐次データ（テキスト、音声）が解析対象とされてきたが、世の中のいろいろな現象、データを対象とすると表現力が不足していた。他方、グラフとは対象（ノード）とその間の関係（エッジ）で表現されるデータ構造と定義される。グラフで表現できるデータは非常に多く、幅広い分野で応用されている。

計算手法の選択と最適な計算条件の設定を行えば、高精度な物性パラメータが得られるものの、分子科学計算とグラフの機械学習の融合が困難であり、分子科学計算（分子軌道計算、分

子動力学)をもとにした記述子は検討が不十分である。この原因の一つは、分子科学計算の結果を機械学習に適したデータ形式に成型することが困難だったことである。

分子科学計算によって得られた分子構造の詳細データを情報の損失なく解析するためには、原子の物理パラメータや化学結合の物理パラメータとして計算結果を取り出すことが必要である。このような原子のベクトル表現、化学結合のベクトル表現を加味した分子構造の表現に対して、グラフ構造による表現は親和性が高い。本検討では 700 種類の匂い分子について行った量子化学計算の結果解析を進めている。

[今後の研究の方向, 課題]

本研究の目指す匂い情報空間の構築は、匂いの質を解析する前提をもたらし、匂いセンシン

グ・情報化・VR 再現におけるインターフェース設計に貢献する。分子科学と機械学習を融合した挑戦は、ケミカルスペースとコンピューターとの調和の促進に貢献する。

匂い情報空間について分子科学計算に基づくケミカルスペース表現において、機械学習アーキテクチャーと分子科学計算を融合させ、設計することが必要である。筆者は量子化学計算と連動したグラフの機械学習(データローダー作成, アーキテクチャー検討)を進める予定である。

[成果の発表, 論文など]

“A study on odor sensing space of odorants chemical space; molecular structural diversity and resolution of descriptors”, Yuki Harada, Shuichi Maeda, Junwei Shen, Taku Misonou, Hirokazu Hori, and Shinichiro Nakamura (投稿中)