《様式Ｂ》

研究テーマ　「化学構造の多様性の可視化に関する研究」

研究責任者　所属機関名　豊橋技術科学大学
官職又は役職　助教
氏　　名　桂樹哲雄　　メールアドレス katsuragi @ cs.tut.ac.jp

共同研究者　所属機関名
官職又は役職
氏　　名

（平成 29 年度募集）第 30 回 助成研究 完了報告書

上記様式記載後

１．実施内容および成果ならびに今後予想される効果の概要（１，０００字程度）

※産業技術として実用化の可能性や特許出願（予定も含む）の有無についてもご記載ください。

地球上では多くの種類の化学物質が用いられているが、それらのすべての安全性が確認されているわけではない。人体や環境への安全性を正確に調べるためには大量の時間とコストがかかる上に、わが国で規制対象となっている 2 万種以上の化学物質のうち、既に安全性が点検済みであるものは約 1600 種のみであり、定量的構造活性相関（Quantitative Structure – Activity Relationship, QSAR）による予測モデルの活用が期待されている。構造が似ている化合物、特に同一の骨格構造を持つ化合物は、しばしば、似た生理活性（医薬品としての作用、毒性など）を持つ。QSAR は、それらの化学構造と生理活性の強弱の間に認められる統計的な関係のことであり、この関係を利用して、構造を手掛かりに生理活性を定量的に算出する方法が QSAR モデリングである。QSAR モデリングにおいては、データベースに既存の化合物の生理活性を登録して参照する。申請者の所属するグループの先行研究により、収録構造の多様性が低く、同一の骨格構造を持つ（すなわち構造が類似する）化合物が他にデータベース内に存在しない場合に、QSAR モデリングの予測精度が悪くなることが明らかになった。そこで、本研究では、安全性が未点検である化学物質の中から次の試験対象物質を選ぶ際に、データベース内に含まれる化学構造がなるべく多様となるように、データベース内に構造が類似する化合物が含まれていないものを見つけるための情報可視化ツールを作成した。天然有機化合物 51247 件の化学構造を用いて、それらが持つ骨格構造に基づいて分類したうえで、自己組織化マップによって似た構造同士を 2 次元マップ上に写像した。さらにこのマップ上に各構造の環数、生理活性などの特徴量を重ね合わせることで、関連付けられた特徴が近いものがよりマップ上のより近い位置に写像されていることがわかった。これらの結果から、データベース内の多様な化合物の構造を構造類似性（多様性）の観点から可視化することができた。

QSAR モデリングにおいて、様々な化合物に対応するためには、既知化合物の構造の多様性が重要である。実際、多様な構造の化合物のデータを得るために、熟練した化学者が経験を頼りに次にどの化合物を選ぶかを判断してきた。本研究の成果を応用すると、デー

タの多様化を目指して化合物の骨格構造データを追加することを考える際に、２次元平面上にマップされた化学構造地図を用いることで、既知の化学構造の分布がどのようなものかを直感的に把握することができるようになる。

２．実施内容および成果の説明（Ａ４で、５ページ以内）

はじめに

自己組織化マップは、T.Kohonen により 1981 年に提案された教師なし学習ニューラルネットワークである。このネットワークは入力層と競合層からなる 2 層のニューラルネットワークで、元の特徴空間における各データの近接関係を保ちながら、より低次元の空間に非線形写像を行うことでデータの分布構造を可視化することができる。

本研究では、自己組織化マップの入力として、データベース内のすべての化学構造から環構造部分を NTG[Takahashi 2004]として取り出し、TFS 法[Takahashi 1998]により得られたパターンベクトルを用いた。このようにすることで、化合物の骨格構造が近い化合物ほど近くに写像された 2 次元化学構造マップを作成することができる。構造多様性は構造類似性の裏返しであり、このマップを参照することで、データベース内の化学構造の多様性を可視化することができる。

分子骨格の NTG による表現

本課題では、NTG を用いて分子骨格を表現する。NTG は、分子の環構造に着目した部分グラフで、次数が 1 以下の頂点原子を持たないグラフとして定義される[Takahashi 2004]。NTG には、頂点原子と結合タイプの種類に重みづけをするか否かで、以下の 4 つの表現レベルを定義することができる。

(1) NTG/SG: 単純グラフ
(2) NTG/VG: 頂点原子の種類の重みつきグラフ
(3) NTG/EG: 結合辺の種類の重みつきグラフ
(4) NTG/CG: 頂点原子、結合辺の種類の重みつきグラフ

NTG は上記の表現レベルに関係なく、環を有する分子グラフに対して、頂点次数が 1 以下の頂点原子を逐次枝狩りをすることで容易に抽出できる。また、上記の４つの表現レベルに加え、NTG に接合する２重結合まで含めた表現レベル（NTG/DG）も定義可能である。

TFS 法

本課題では、化学構造の特徴量を数値ベクトルとして扱うために、TFS 法を用いる。TFS[Takahashi 1998]は、高橋らによって考案された構造情報を特徴づけるための数値的な記述手法の一つである。TFS は、(1) 対象の構造について、可能なすべての部分構造フラグメントを生成・列挙し、(2) それぞれの部分構造フラグメントに対して、各頂点の次数などを用いて数値的特徴づけを行うことで生成できる。この特徴量の出現頻度のヒストグラムが TFS であり、構造情報を多次元の数値ベクトルとして扱うことができる

自己組織化マップ

本課題では、高次元データ空間に浮かぶ化学構造同士の近接関係を平面に非線形写像することで可視化することを狙いとして、自己組織化マップ(SOM)を用いる。SOM は入力層と競合層からなる 2 層ニューラルネットワークで、元の特徴空間における各データの近接関係を保ちながら、より低次元の空間に非線形写像を行うことでデータの分布構造を可視化することができる（図１）。

TFS 法によって得られた化学構造特徴量ベクトルを入力とした SOM の場合、化学構造の特徴を表現した TFS が類似していると、位置的（位相的）に近くに写像される。これは、SOM が非線形写像であることに由来する。これにより、スパースな領域が折りたたまれて、化学構造の類似関係が表現される。

入力層には、１回の SOM 学習ごとに、すべての高次元ベクトルを入力する。各入力ベクトルは、学習 1 回ごとに最も類似する重みをもつニューロン(BMU)を探し出し、近傍ニューロンの重みが入力ベクトルにさらに近づくよう、重みを修正する。このように入力ベクトルとニューロンの重みベクトルを比べながら学習するため、競合層に並ぶニューロンのそれぞれの次元数は入力ベクトルの次元数と同じである。

本課題では、初期マップの生成の際に、主成分分析を用いる線形な初期化を行う。すなわち、最も大きな固有値を持つ自己相関行列の２つの固有ベクトルを決定し、次にこれらの固有ベクトルを２次元の線形部分空間に張り付ける。長方形配列をこの部分空間に沿って、その重心がデータセットの平均と一致するように設定する。

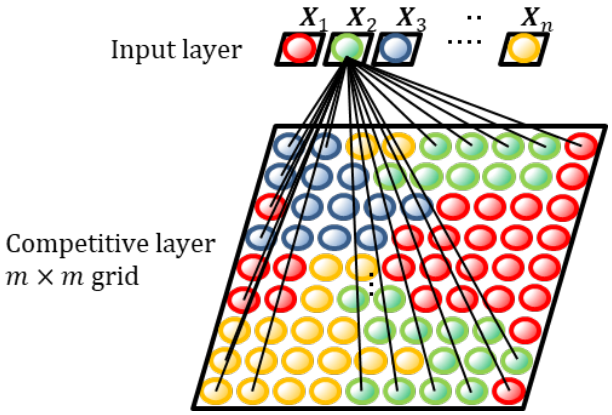

図 1 SOM の学習の概要

データセット

本課題では奈良先端大で開発された天然有機化合物データベース KNApSAcK Family [Afendi et al. 2012]に収録されている化合物の構造データ 51247 件および生理活性データ[Nakamura et al. 2014]に収録されている化合物-生理活性情報 4426 対を用いた.

実験

KNApSAcK Family データベース内の全化学構造から骨格構造を抽出した。その際に、原子数が 85 以上の分子は取り除いた。得られた骨格構造から TFS を生成し、TFS に基づいて SOM の学習を行った。なお、マップサイズは 10x10, ベクトル更新法を用いて、10000

エポックの学習を行った。得られたマップについて、構造地図、環数、生理活性の観点から解析を行った。以下に、結果を示す。

まず、全化合物から得られた NTG/SG に対する結果を示す。図 2 に、SOM 上の 10x10 のセル内に配置されたそれぞれの化合物から抽出した NTG/SG をセル毎にらせん状に並べ、その環数に応じて色分けした結果を示す。図 2 の結果から、1 から 18 環系の化学構造がデータベース内に収録されていることが可視化された。それぞれのセル内に、近い環数の構造が配置されていること、また、近傍のセルにも同様に近い環数の構造が配置されていることがわかる。化学構造が似ていると、その環数は同じか近い値となるはずであり、構造の類似性を簡易的に可視化する指標と考えられ、この結果から、より類似した化学構造が、近傍のセルに配置されていることが確認できた。また、データベース内の化学構造の多様性について、環数の観点から可視化することができた。

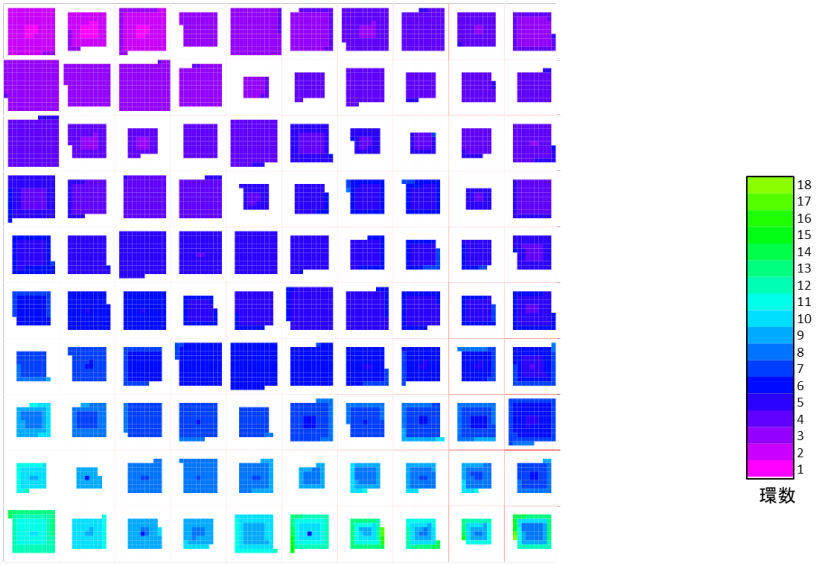

図 2 データベース内の全 NTG/SG を対象とした SOM（環数で色分け）

図 3 には、同じく全化合物から得られた NTG/SG を対象とした場合の、それぞれの NTG/SG に関連付けられた生理活性の種類で色分けした SOM を示す。生理活性の偏りは少なく、SOM 全体にわたって分布していることがわかる。構造が似ている化合物は似た活性をもつことが多いため、SOM 上での生理活性の分布には偏りが生じることが期待されるが、そうはなっていないことがわかった。これは、NTG/SG を用いた解析では頂点原子の情報と結合の種類を考慮しなかったために、分解能が足りなかったためと考えられる。

次に、さらに生理活性と構造の関係をさらに詳しくみるために、3,4 環系の NTG/SG のみを対象とした解析を行った結果を図 4 に示す。この場合にも、全化合物を対象にした場合と同様、それぞれの生理活性の分布に偏りはみられない。これも、上記と同様に NTG/SG の分解能の不足が問題となったと考えられる。この点は NTG/CG 等を用いた解析を行えば解決すると考えられる。

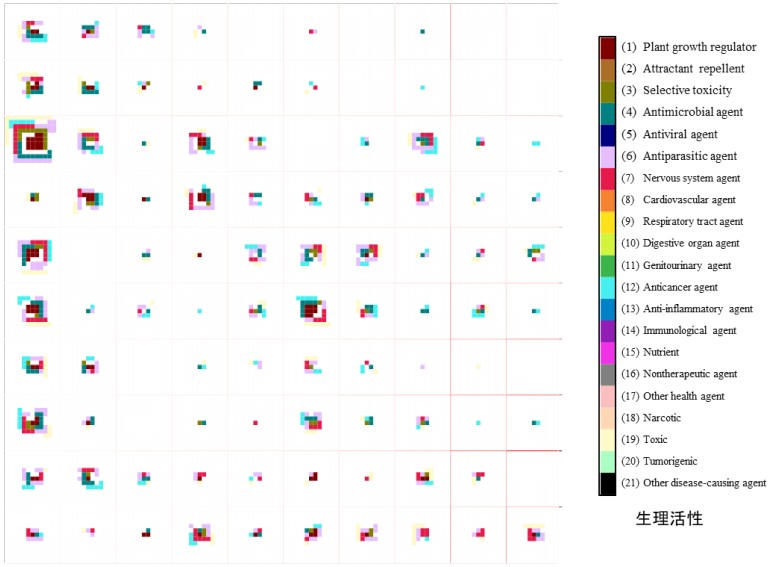

図 3 データベース内の全 NTG/SG を対象とした SOM（関連する生理活性で色分け）

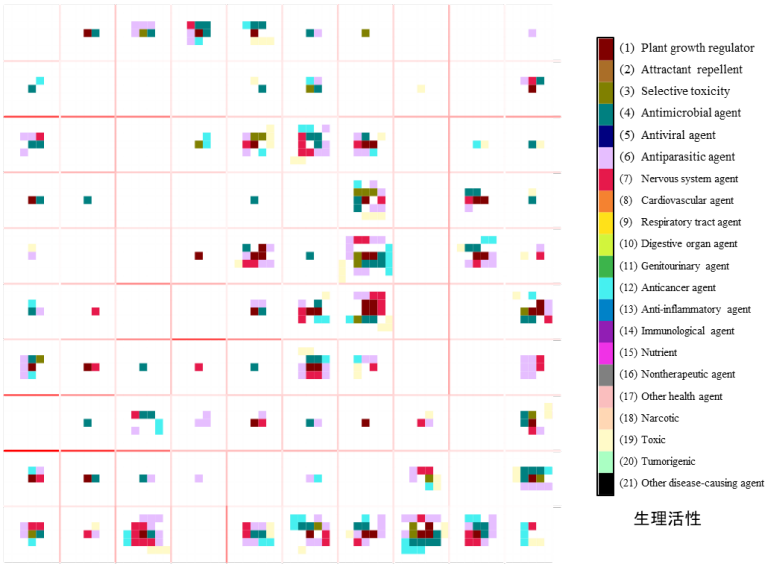

図 4 データベース内の 3,4 環系 NTG/SG のみを対象とした SOM
（関連する生理活性で色分け）

参考文献

[Afendi et al. 2012] F. M. Afendi et al., Plant Cell Physiol., 53, e1 (2012).

[Nakamura et al. 2014] Y. Nakamura *et al.*, *Plant Cell Physiol.*, **55**, e7 (2014).

[Takahashi 1998] Y. Takahashi, H. Ohoka, Y. Ishiyama, in Advances in Molecular Similarity, vol. 2, (R. Carbó-Dorca and P. G. Mezey, Eds.), pp. 93-104, 1998.

[Takahashi 2004] Y.Takahashi, In 2004 IEEE Intl. Conf. on Systems, Man and Cybernetics, vol. 5, pp 4583–4587, 2004.

関連成果（国際会議発表）

[1] T. Katsuragi, Y. Takahashi, “NTG-Activity map of natural organic compounds based on self-organizing map,” 22nd EuroQSAR, PP132, Thessaloniki, Greece, Sep. 16-20, 2018.

[2] T. Katsuragi, Y. Takahashi, “NTG-activity dictionary of bioactive natural organic compounds and structure similarity search”, The 11th Japan-China Joint Symposium on Drug Discovery and Development, Shaoxing, China, Jun. 24, 2018.