2G04

第一原理計算を援用したベイジアン・アプローチによる新規化合物探索

（北陸先端大[1]、物材機構[2]、総研大[3]、地球快適化インスティテュート[4]、統数研[5]）
○本郷研太[1,2]、 池端久貴[3]、 磯村哲[4]、 前園涼[1]、吉田亮[2,3,5]

A new Bayesian approach to chemical compound search associated with ab initio simulations

(JAIST[1], NIMS[2], SOKENDAI[3], The KAITEKI Institute, Inc.[4], ISM[5])
○Kenta Hongo[1,2], Hisaki Ikebata[3], Tetsu Isomura[4],
Ryo Maezono[1], and Ryo Yoshida[2,3,5]

【緒言】

ケモインフォマティクス[1]の究極目標は、「所望の物性 Y を実現する化合物構造 S の予測」にある。この問題は、構造 S から物性 Y を予測する定量的構造物性相関（QSPR; Quantitative Structure-Property relationship）を順問題と捉えれば、その逆問題 Inverse-QSPR とみなせるが、その最適解を求めるには、探索空間は余りにも広大すぎる：既知の化合物数は 10^9 程度であるのに対して、化合物空間には 10^{62} 程度の化合物が存在すると推測されている[2]。先進的な方法論が多数報告されている QSPR 研究とは対照的に、Inverse-QSPR 研究は限定的で、グラフ数え上げ[3]や遺伝的アルゴリズム[4]を基づき、化合物フラグメントから新規化合物を生成・探索する先行事例が報告されているのみである。しかしながら、フラグメント法で効率的かつ大域的探索を実現するには、フラグメントの種種や総数に依存して、探索空間の制限や、分子改変手続きにおける計算量増大の問題など、いくつかの課題が残されている。

我々は最近、順問題（QSPR モデル）に第一原理計算を援用し、「順問題の知見獲得を活かした逆問題解法」であるベイズ統計を探索指針とする、新しい化合物構造探索手法を開発し、これまでに、有機分子系での概念実証に成功している[5]。本手法は、化合物生成に自然言語処理モデルを適用することで、従来手法における探索空間と計算効率の問題を回避している。また、従来のケモインフォマティクスのアプローチとは異なり、実験結果の存在しない、「未知化合物」を提案するため、その物性計算に計算科学を援用している。すなわち、急激に実用性を増す「計算科学」と「情報科学・統計科学・データ科学」の融合展開にも、本手法の新規性を見いだせる。

【方法論】

本研究の問題設定は、「所望の物性域 $Y \in U$ に属する化合物構造 S の探索」であり、有機分子系を対象に、それらの HOMO-LUMO ギャップと内部エネルギー値を対象物性とした。本研究のベイジアンアプローチでは、条件付き確率分布（事後分布） $p(S|Y \in U)$を導入し、$p(S|Y \in U)$の高確率領域に分布する構造 S をモンテカルロ法で探索していく。この $p(S|Y \in U)$は、ベイズの定理 $p(S|Y \in U) \propto p(Y \in U|S)\, p(S)$ より、「尤度（順問題の QSPR モデル）」$p(Y \in U|S)$と「化合物構造に関する事前情報/分布（化合物らしさ）」$p(S)$の積として得られる。本研究の $p(Y \in U|S)$は、線形回帰モデルを採用し、PubChem データベースからランダムに選んだ 10,000/6,674 個の化合物物性を学習/テストデータとして QSPR モデルを構築した。物性計算は、Gaussian09 を用いて、密度汎関数法（DFT/B3LYP）により算出した。本研究で開発した化合物構造生成法 $p(S)$は、化合物を SMIILLES 形式の文字列とみなし、化合物表現用に拡

張した自然言語処理モデルを適用し、文字列の出現頻度を PubChem データベース収録の 50,000 化合物構造データを用いて学習した（計算手法・条件の詳細は文献[5]に報告）。

【結果と考察】

本研究は、内部エネルギー（kcal/mol）と HOMO-LUMO ギャップ（eV）をターゲットとして、所望の物性量を持つ化合物を探索していく。物性探索の値域 U として、U_1 = [100, 200]×[4, 5.5]、U_2 = [100, 250]×[2.5, 3.5]、U_3 = [250, 400]×[5, 6]の 3 領域を設定した（それぞれ、図 1 中の赤、青、緑のハッチング部分）。フェノール分子を初期構造として、各物性域を目指して、化合物を探索していく。図 1 は、モンテカルロステップ t = 20/50/200 毎の物性量プロットと候補化合物である。上図の丸印は、候補化合物のうち、尤度最大の上位１０化合物を示しており、その上位４化合物の分子構造を下図に示している。t = 20/50 では、初期分子構造であるフェノール分子のベンゼン環構造を踏襲した分子構造が出現しているが、t = 200 まで到達すると、ベンゼン環構造は消失している。このことは、本研究の分子生成手法が多様な化合物構造を生成し得る可能性を示唆している。

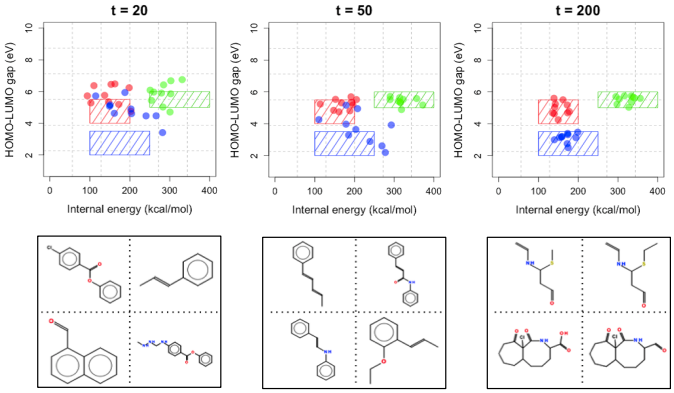

図 1：（上段）モンテカルロステップ t = 20/50/200 における生成化合物の物性値分布。各物性域探索 U_1/U_2/U_3 につき、尤度最大の上位 10 化合物を表示している。（下段）U_1 に対して、各 t で尤度最大の上位４化合物の分子構造。

図 2 (a)/(b)/(c)はそれぞれ、t = 500 における、物性域 U_1/U_2/U_3 に属する化合物である。同図上段は、各物性域に属する化合物候補として推測された分子構造であり、下段は PubChem データベースに実在する、上段化合物との類似性が極めて高い分子構造（Tanimoto 係数 > 0.9）を示している（各物性域のパネル左）。各物性域のパネル右は、データベース上にも存在しない新規性の高い「埋蔵分子」を発見した可能性が極めて高いことを示唆している。

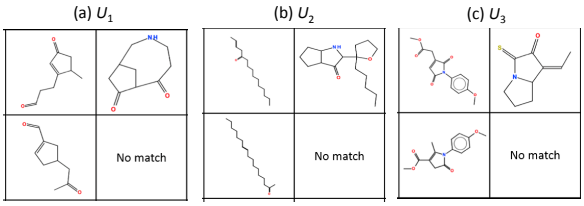

図 2：（上段）当該探索アルゴリズムが発見した各物性域に属すと推測される化合物候補。（下段）上段分子と極めて類似性の高い PubChem 登録分子。

本講演では、このベイジアンアプローチに基づく化合物構造探索の詳細を報告するとともに、その汎用性について議論する。

【参考文献】

[1] Gasteiger, J. (ed.); Engel, T. (ed.), "*Chemoinfomatics*: *A Textbook*", John Wiley & Sons, 2004.

[2] Lahana, R. *Drug Discovery Today*, **1994**, *4*, 447-448.

[3] Miyao, T.; Hiromasa, K.; Funatsu, K. *J. Chem. Inf. Model.* **2016**, *56*, 286-299.

[4] Wong, W. W.; Burukowski, F. J. *J. Cheminform.* **2009**, *1*, 4.

[5] Ikebata, H.; Hongo, K.; Isomura, T. ; Maezono, R.; Yoshida, R. (投稿準備中).