De novo デザインのための ring systems に基づいた化学構造創出

宮尾 知幸, 金子 弘昌, 船津 公人*

Ring-system-based Chemical Structure Enumeration for *de Novo* Design

Tomoyuki Miyao, Hiromasa Kaneko, and Kimito Funatsu*

Department of Chemical System Engineering, School of Engineering, The University of Tokyo; Faculty of Engineering Bldg. 3, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan.

(Received August 16, 2015)

De novo molecular design aims to propose molecules exhibiting desired properties and/or activities, as constructed from scratch. Although this approach opposes the widely used virtual screening (VS), the same criteria should be applied, such as ones based on substructure filters, and quantitative structure-property relationship (QSPR) and quantitative structure-activity relationship (QSAR) regression models. QSPR/QSAR, which enables us to predict properties/activities by making use of experimental data, are widely used in academia as well as in industry. Herewith, we present a novel chemical structure generation system by combining fragments whose final chemical structures satisfy the aforementioned criteria. Using inverse analysis, QSPR/QSAR models determine a specific region in chemical space corresponding to a set of desired values by a designer. Chemical structures are generated by combining ring systems, as well as atom fragments, in every possible way until violating the upper bounds of that region. We also show the results of inverse-QSAR analysis for the human Alpha-2A adrenergic receptor. This suggests that our system has features preferable to VS-like methods in terms of the number of generated structures.

Key words—*de novo* design; ring system; ligand-based design; structure generator

1. 背景

創薬開発の初期段階において virtual screening (VS) を利用し、スクリーニング対象となる自社又は商用ライブラリからメディシナルケミストにとって不要な構造を削除する手法はここ 10 年間にあたり前のこととして認識されるようになった。この際に quantitative structure-activity relationship (QSAR) による活性予測モデル、部分構造検索によるタブーリスト、quantitative structure-property relationship (QSPR) に基づく物性予測モデル、などが適用される (VS におけるフィルタの種類や特性については文献参照のこと)。¹⁾ リガンドが間接的に標的マクロ分子の情報を保有すると仮定した QSAR モデルや QSPR モデルは回帰モデルであるために計算速

度が速く、何百万という構造を含む大規模ライブラリに対しても適用可能である。結果として標的マクロ分子に対して予測活性の高い drug-like な分子構造を得ることができるため、VS を分子設計手法とみなすこともできる。この方法は、薬剤候補化合物以外の機能性分子の設計にも適用することができ、様々な特性が同時に要求される分子を設計 (提案) する際には、目的に合わせた複数のフィルタを化合物ライブラリに適用することになる。この分子設計手法の特徴は、適用対象のライブラリとスクリーニング手法が独立していることである。ライブラリは仮想的なものであっても、実在の化合物構造を収めたライブラリでも問題なく、スクリーニング手法は、ライブラリと独立して開発することになる。

一方で *de novo* デザインとは、分子を最初からデザインすることを意味する。したがって、提案構造は既往のライブラリに含まれる構造に限定されることはない。リガンドに基づく *de novo* デザインによる分子設計のフローをリガンドに基づく VS による分子設計フローと対比させる形で Fig. 1 に示

東京大学大学院工学系研究科化学システム工学専攻 (〒113-8656 東京都文京区本郷 7-3-1 東京大学工学部 3 号館)

*e-mail: funatsu@chemsys.t.u-tokyo.ac.jp

本総説は、日本薬学会第 135 年会シンポジウム S66 で発表した内容を中心に記述したものである。

Fig. 1. Comparison of Virtual Screening (VS) Workflow and Corresponding *de Novo* Drug Design Workflow
Left: VS, right: *de novo*.

す。複数のモデルが適用可能と判断される範囲内で、設計者が期待する性質を持つ分子を設計することが両手法の目的となる。*De novo* デザインでは、目的となる物性・活性値を入力することでコンピュータがこれらの値を満たす分子構造を創出する。この際には、QSPR モデル、QSAR モデル、そして上記に記載したフィルタなどが VS とは反対方向に適用されることになる。あるコンピュータで扱えるライブラリの構造数に制限があると仮定する。例えばそれを 100 万構造とした場合、VS では 100 万構造から望ましい構造を選択することから、結果としてライブラリには不要な構造が存在することになる。一方で、*de novo* デザインでは創出した 100 万構造すべてが望ましい構造となる。このように *de novo* デザインは VS と比較して分かるように、分子設計を効率化できる。

De novo デザインにより分子を設計する手法としては、遺伝的アルゴリズムを利用した方法²⁾が広く研究されている。この方法では、QSPR/QSAR モデルを評価関数として組み入れることにより、予測活性が高い構造のみを提案することができる。構造を網羅的に組み立てるのではなく、初期構造を改変させ、評価関数による評価を行い、高い評価値を持つ構造のみを次世代へ渡す。この際、構造の数を一定に制限することが通常の操作となる。また、構造を改変させる際に、合成可能性を織り込む研究もされている。本稿では、遺伝的アルゴリズムを利用す

る *de novo* デザインではなく（合成可能性も考慮しない）、inverse-QSPR/QSAR を利用したデザインを紹介する。この手法の特徴は、QSPR/QSAR モデルを反対方向に解析することである。目的となる物性あるいは活性： y の値をモデルに入力することで、対応する分子の抽象的表現である記述子： x の値を取得する。次に、構造生成機を用いてこの x の値に対応する構造のみを生成する。遺伝的アルゴリズムを利用した手法とは異なり、 x の情報に焦点を当てて構造生成を行うため、 x の値が目的の y の値に対応する範囲に存在すると仮定すれば、望ましい物性あるいは活性値を持つすべての構造が提案されると期待できる。

2. Inverse-QSPR/QSAR

Inverse-QSPR/QSAR は概念としては単純であるにもかかわらず、研究開発はそれほど行われていない。Inverse-QSPR の例として、Skvortsova は Kier インデックスの特定の値に対応する頂点次数と辺の組み合わせを取得する手法を提案した。³⁾ ケーススタディとして、蒸発熱を目的物性とし、線形重回帰モデル (multiple linear regression; MLR) を利用した inverse-QSPR 解析を実施し、特定の化学構造群の生成に成功している。Inverse-QSPR/QSAR 解析を利用した分子設計が技術的に難しい理由を以下の 4 点にまとめる。(1) QSPR/QSAR モデルの写像関係 ($x \rightarrow y$) が全単射ではないこと、(2) 化学構造と x との関係が全単射ではないこと、(3) 化学構造を生成する際に、構造生成機が必要となること、(4) inverse-QSPR/QSAR 解析を実施するためには、入力目的物性→逆解析→記述子情報→構造生成というプロセスを 1 つのシステムとして構築する必要があること、である。(4) について詳しく述べる。Inverse-QSPR/QSAR 解析を行うためには、特定の種類の記述子を使用する必要がある。例えば、radial basis function 記述子や highest occupied molecular orbital 及び lowest unoccupied molecular orbital のエネルギー値を記述子として使用した場合、記述子の値が特定の値をとる構造を生成することは（原理的には可能であったとしても）難しい。また、分子量という単純な記述子を考えた場合でも、同一の分子量を持つ構造は多く存在し、分子量という特性に合わせた構造生成機が必要となる。つまり、記述子から構造を生成するためには、構造生成機は指定した記述

子に特化したものとならざるを得ず、汎用的な構造生成機を用いることは効率性の観点から避けるべきである。したがって、指定した記述子に特化した機能を構造生成機に導入しようとした場合、構造生成機と記述子とはもはや切り離すことのできない関係となる。このような記述子と構造生成機とが一体となった inverse-QSPR/QSAR 研究の一例として、化学グラフにおける特定の種類のパスの総数を記述子として用いて、構造生成機においてパスの総数に対応する化学構造を生成する研究などが行われている。⁴⁾ したがって一般に自由な記述子を利用して inverse-QSPR/QSAR 解析を行うことはできず、構造生成機に組み込むことが可能な記述子という制約の中で、分子の抽象的表現である記述子をどのように採用すべきかという命題を考える必要がある。

3. 提案する inverse-QSPR/QSAR 構造生成システム

分子構造に対して部分構造（フラグメント）を付け加えることにより記述子の値が単調に変化する記述子を使用した場合、提案構造の網羅性を失うことなく、効率的に構造を生成することができることをわれわれは示した。⁵⁾ このような性質を持つ記述子の総称を *monotonous changing descriptors* (MCDs) と名づけた。MCDs は分子量や特定の部分構造の数といったカウントに基づく記述子から、Randic 結合インデックス、自己相関記述子といった分子全体の特徴を表現する記述子まで含めることができ、MCDs を用いることで予測精度の高いモデル構築が基本的に可能になると考える。MCDs は分子量と高い正の相関関係にある記述子のみならず（多くの MCDs は正の相関を持つが）、構造生成する経路に応じて記述子の値が増加すれば MCDs と分類されるので少なくとも強い相関を示すとは限らない。実際にあるケーススタディでは、とりわけ部分構造の数などの記述子は分子量と低い相関を示す結果となった。⁶⁾ 提案システムでは MCDs を QSPR/QSAR モデルを構築するために用いる記述子としている。MCDs を用いた記述子空間（MCD 空間）において逆解析を行うことで、MCD 空間内に特定の領域を定めることができる。規定した領域に存在すると考えられる構造を、部分構造（フラグメント）を組み合わせることにより組み上げる。Ring systems (RSs)⁷⁾ と一種類の原子価則（結合手の数）に

Fig. 2. An Example of Trajectory of Generating Structures in Monotonous Changing Descriptors (MCDs) Space

Values of MCDs necessarily increase by adding fragments. Once one of these values exceeds the upper bound of MCDs, pruning is conducted. Dots are the coordinates in MCDs space with the corresponding structures on their right-side. Since anisole exceeds the upper bounds of MCD2, the structure is discarded.

基づいた原子フラグメントを構造生成の部品として、それらを化学グラフとしてあらゆる方法で組み合わせる。制約条件の上限を超えた時点で成長構造の枝刈りを実施するが、MCDs を利用して領域を規定しているため、枝刈りにより構造の成長を終了させた場合であっても制約条件を満足する構造の網羅性は失われない (Fig. 2)。

一般に QSPR/QSAR モデルを利用して物性・活性値の予測をする場合には、applicability domain (AD: 適用範囲) を考慮する必要があるが、本システムにおいて逆解析を行う際も例外ではない。AD を考慮しなければ、回帰モデルの外挿が際限なく許されることとなるが、そのような外挿領域における構造の物性・活性予測値の信頼性は当然ながら低い。AD モデルとして、回帰モデル構築に利用したデータの密度や分布を利用する手法が提案されている。⁸⁾ 本研究においては、狭義の AD としてトレーニングデータの事前分布 $[p(\mathbf{x})]$ の密度の高い領域を利用することとした。狭義の AD とは回帰モデル構築に利用する記述子空間におけるデータの高密度領域を意味する。化学構造を記述子に変換し、回帰モデルを適用する QSPR/QSAR 解析の場合には、予測対象となる化学構造とモデル構築に用いた化学構造との類似性のある程度保つ必要もある。この主張の妥当性は直感的には明らかであるが、実際に化学構造同士の類似性を表現する方法は多種多様であり、一意に類似性を規定することはできない。

こうした類似性を構造生成により生成した構造とモデル構築に用いたデータとの間で考慮するために、構造を生成するための部分構造(フラグメント)は、トレーニングデータに含まれる化学構造から抽出した RSs と一種類の原子価則に基づいた原子フラグメントを用いることとした。これらの部分構造(フラグメント)を組み合わせることで構造を生成する。

下記に AD を考慮して inverse-QSPR/QSAR 解析を行う手法として、ベイズの定理を利用して目的変数: y の値に対応する記述子: x の情報を密度分布として表現し、MCD 空間において密度が高い領域を定める方法について述べる。先に概説したように逆解析においては、回帰モデルを利用することから AD 内に含まれる記述子の値を構造生成の対象とすべきである。混合正規分布による事前分布: $p(x)$ の密度が高い領域を AD と定めるが、inverse-QSPR/QSAR 解析の結果に応じて、目的となる y の値に対応する x の密度関数は変化することに注意を要する。QSPR/QSAR モデルとして最小 2 乗法による MLR モデルを用いた。最小 2 乗法による MLR モデルは平均を回帰直線、分散をガウス誤差とした正規分布として表現することができる [$p(y|x)$]。事前分布 $p(x)$ と組み合わせることで y の値を所与とした際の x の事後密度分布 $p(x|y)$ を同じく混合正規分布として表現できる。⁵⁾ 事前分布により AD を考慮した逆解析が可能となる。逆解析により y に対応する x を写像関係として規定することはできないが、確率密度として x の情報を規定することで、1 つの y に複数の x が対応する状況を表現することとなる。最後に $p(x|y)$ の密度関数の密度が高い領域を構造生成のための制約条件とする。

4. 構造生成機

部分構造(フラグメント)を組み合わせることにより構造生成を行う手法(プログラム)は古くから存在する。⁹⁾ このタイプの構造生成機では、単純に部分構造(フラグメント)を組み合わせることだけを考えたために、当然のこととして組み合わせ論的爆発に陥る。化学空間の大きさは 10^{60} – 10^{200} と推定されているが、^{10,11)} これらの推定値は、化学構造を化学グラフとみなし適当なサイズの分子を想定した場合に、組み合わせ論を用いて原子価則に違反しない場合にどの程度の数の構造が許容されるのかを計算した結果に基づくものが多い。Figure 3 に 289

Fig. 3. The Number of Estimated Structures When Combining Fragments

Fragments are 289 RSs and C, N, O, F, Cl, Br, I atom fragments. How to estimate such number is described in the text body.

個の RSs と C, N, O, F, Cl, Br, I を原子フラグメントとして利用した場合に、組み合わせる部分構造(フラグメント)の数と対応する構造数を示す。構造数の見積もりは文献に記載された方法¹²⁾を利用した。部分構造(フラグメント)を付け加えることで、新たな構造が生成するため、付け加える前の構造と付け加えた後の構造とは親と子の関係があるとみなせる。生成するすべての構造においてこの関係を明示した場合に、構造同士のつながりは、1 つの部分構造(フラグメント)を頂点とした木構造(生成木)となる。生成木において深さに応じた特定の確率で枝刈りを実施することにより、枝刈りを行わなかった場合に発生すると想定される構造数を計算できる。Figure 3 では約 300 種類の部分構造(フラグメント)を部品とし、それらから重複を許して 10 個を組み合わせると 6×10^{21} 個の構造が生成している。したがって構造を網羅的に組み合わせ、その後不要な構造を削除するという方針で構造を生成することは現実的ではないことが分かる。

われわれの提案手法の前提となる仮説は、「MCD 空間において小さく規定した領域に存在する妥当な化学構造の数は、組み合わせ論的爆発を抑える形で構造生成が可能な数である」という考えである。提案手法の目的はそのような構造を生成することにある。妥当な化学構造とは、熱力学的・速度論的に不安定な構造、reactive な構造を除外した構造に加え、モデルの適用範囲に存在する構造を意味する。また、網羅的な構造生成が現実的な時間において行うことができない場合(不十分な領域の規定又は仮

説の棄却の場合)に、規定した領域内に存在する、できるだけ多様な分子構造を生成するシステムを構築することが目標となる。今回、上記のように様々な制約条件を考慮した場合であっても、漏れや重複がなく構造を生成するためにカノニカル経路生成法¹²⁾を RSs に適用可能な形に改良し、実装を行った。今回この仮説の検証に向けての取り組みの一例として、inverse-QSAR 解析の結果を示す。

5. ケーススタディ⁶⁾

GVK データベース¹³⁾より、alpha-2A adrenergic receptor (ADRA2A) に対する阻害定数の逆数の対数値 pK_i を持つ 643 分子を取得し、inverse-QSAR 解析を実施した。ランダムに抽出した 500 分子を QSAR モデル構築のためのトレーニングデータに、残りの 143 分子をモデルのバリデーションのためのテストデータに分類した。MCD として分子量、結合数、重み付き多重結合の数、水素を除く原子の数、水素結合供与体の数、環の数、5 員環の数、1st Zagreb インデックス、Randic 結合インデックス、窒素原子とフッ素原子のトポロジカル距離の総和、部分構造の数 [ArCO, ArCN, ArNR₂, RCH₃, XCH₃, R₂CH₂, CHR₃, ROH, =O, XH (Ar: アロマトニックな原子, R: アロマトニックではない炭素, X: ヘテロ原子)] の計 20 個を用いた。

モデルの適用範囲を示す $p(\mathbf{x})$ は正規分布の組み合わせにより表現した。各正規分布の平均ベクトル及び共分散行列は EM アルゴリズムにより決定した。また、MLR モデルによる QSAR モデル構築の結果、 $R^2=0.78$, $RMSE=0.44$ となり、テストデータに対するモデルの精度は、 $R^2_{pred}=0.74$, $RMSE_{pred}=0.46$ となった。この結果を先ほど計算した $p(\mathbf{x})$ と組み合わせることで、 $p(\mathbf{x}|\mathbf{y})$ を計算し、MCD 空間上に構造生成するための領域を規定した。今回のケーススタディでは望ましい活性の値を 9 とし $\mathbf{y}=9$ の条件を与えた際の $\mathbf{x}$ の事後分布を導出し、導出した同時確率密度分布の各 MCD に対する周辺尤度分布の確率が 40% となる領域に上限と下限を定め、構造生成のための制約条件とした。構造生成の部品として、トレーニングデータから抽出した 193 個の RSs、原子フラグメントとして C, N, O, F を用いた。また、構造生成の際には、C=C, RSs 同士の直接結合禁止、ヘテロ原子同士の直接結合を禁止するルールを加えた。計算機の性能は、Windows 7

Table 1. Results of Structure Generation Based on 40% of Probability for Each MCD⁶⁾

Inverse QSAR method			VS-like method		
Fragments ^a	Time [s]	StrsOK ^b	Time [s]	Strs ^c	StrsOK ^b
3	0.01	0	0.04	5054	0
4	0.08	0	0.81	121800	0
5	0.80	23	24.28	3386826	23
6	7.99	277	715.16	103925780	277
7	80.57	1656			
8	786.15	36580			
9	7617.00	416269			

The left part is the proposed method, where the right one corresponds to the results without using MCD constraints (the number of necessary structures when applying VS-like methods). ^a number of fragments combined, ^b number of structures inside the ranges of constraints, ^c number of generated structures, cut-off time is 7200 [s].

PC, CPU は 3.33 GHz Intel Xeon, RAM は 16 GB である。構造生成の結果を Table 1 にまとめる。Inverse-QSAR 解析により提案される構造の数と、仮に QSAR モデルをフィルタとして使用した場合に必要な数とを比較する。Inverse-QSAR 解析を利用した構造生成では、構造の数が VS を模倣した手法と比較して少なくなっているのは当然であるが (MCDs による制約条件を加えているため)、構造生成数の増加幅が指数関数的増加と比較して緩やかになっていることが分かる。この結果は、われわれの仮説を実証するものではないが、MCD による制約を利用することで構造生成数が組み合わせ論的爆発を起こさない可能性を示唆するものとなった。生成した総構造の予測 pK_i を QSAR モデルにより計算した結果を Fig. 4 に示す。生成構造の予測 pK_i が目的とする値である 9 に近いことがわかる。この結果から、inverse-QSAR により目的とする活性のみに焦点を当てた構造生成が可能になったことが分かる。

6. 結論

リガンドベースの *de novo* デザイン手法として、新規構造を効率的に創出する可能性のある inverse-QSAR/QSPR を利用した構造生成について紹介した。この inverse-QSAR/QSPR 解析方法では、確率密度をベイズの定理を利用して組み合わせることを提案している。ガウス分布は分布の組み合わせに対して解析解を導出できるため、混合正規分布としてモデルの適用範囲を指定し、ガウス誤差を仮定した MLR による回帰モデルを組み合わせることで、

Fig. 4. Expected pK_i Based on the QSAR Model Used for Inverse-QSAR Analysis⁶⁾

$p(x|y)$ の導出が可能となる。導出された密度関数から、特定の確率となる領域を指定し構造生成を実施する。フラグメントを組み合わせることで構造を生成するにあたっては、組み合わせ論的爆発を抑制することがまず求められるが、われわれはMCDによる制約条件を厳しくすることに加え、1つのRSを1つの頂点と捉えるアルゴリズムの導入により対処することを試みている。提案手法はADRA2Aのリガンドデザインに対して適用され、構造の数が組み合わせ論的爆発に陥っていない可能性を示唆する結果とともに、inverse-QSARを利用することで活性が望ましい値のみに焦点を当てた構造生成が可能となった結果が示された。

謝辞 本研究の一部は、科学技術振興機構(JST)の戦略的創造研究推進事業(CREST)、研究領域「科学的発見・社会的課題解決に向けた各分野のビッグデータ利活用推進のための次世代アプリ

ケーション技術の創出・高度化」, 研究課題名「医薬品創薬から製造までのビッグデータからの知識創出基盤の確立」の支援により行われた。ここで感謝の意を示す。

利益相反 開示すべき利益相反はない。

REFERENCES

- 1) Walters W. P., Stahl M. T., Murcko M. A., *Drug Discov. Today*, **3**, 160–178 (1998).
- 2) Mishima K., Kaneko H., Funatsu K., *Mol. Inf.*, **33**, 779–789 (2014).
- 3) Skvortsova M. I., Baskin I. I., Slovokhotova O. L., Palyulin V. A., Zefirov N. S., *J. Chem. Inf. Comput. Sci.*, **33**, 630–634 (1993).
- 4) Churchwell C. J., Rintoul M. D., Martin S., Visco D. P., Kotu A., Larson R. S., Sillerud L. O., Brown D. C., Faulon J. L., *J. Mol. Graphics Model.*, **22**, 263–273 (2004).
- 5) Miyao T., Arakawa M., Funatsu K., *Mol. Inf.*, **29**, 111–125 (2010).
- 6) Miyao T., Kaneko H., Funatsu K., *Mol. Inf.*, **33**, 764–778 (2014).
- 7) Bemis G. W., Murcko M. A., *J. Med. Chem.*, **39**, 2887–2893 (1996).
- 8) Kaneko H., Funatsu K., *AIChE J.*, **59**, 2046–2050 (2013).
- 9) Rarey M., Stahl M., *J. Comput. Aided Mol. Des.*, **15**, 497–520 (2001).
- 10) Bohacek R. S., McMartin C., Guida W. C., *Med. Res. Rev.*, **16**, 3–50 (1996).
- 11) Kirkpatrick P., Ellis C., *Nature*, **432**, 823–823 (2004).
- 12) McKay B. D., *J. Algorithms*, **26**, 306–324 (1998).
- 13) GVK database: <http://www.gvkbio.com>, cited 25 July, 2015.