G-003

モンテカルロ木探索による特定の有機低分子化合物を出発点とした誘導体の生成手法の開発
Molecular Generation from specific molecules: Extension of ChemTS

恵利川 大樹† 安尾 信明‡ 関嶋 政和†,‡
Daiki Erikawa Nobuaki Yasuo Masakazu Sekijima

1. はじめに

新しい医薬品を開発するためには平均 13.5 年かかり、開発費用は平均 26 億ドルといわれている[1]。そこで情報技術を活用することによって、この膨大な期間と費用の削減が期待されている。創薬のプロセスの一つに化合物最適化があり[2]、特定の化合物を出発点としてより望ましい物性を持った化合物を探索するということが行われている。

創薬において考えられる化合物の数は 10 の 60 乗のオーダーといわれており[3]、その中から望ましい物性を持つ化合物を探索しなければならない。近年の計算機の発展により、大規模なデータを高速に処理することが現実的になったことを受けて、データ駆動型の化合物生成手法として機械学習を利用した生成モデルが注目されている[4]。画像生成などの他分野で成功した手法の多くが利用され、現在までに一から化合物を生成する様々な機械学習手法が提案されてきた[5][6][7][8]。

本研究では従来から行われてきた一から化合物を生成するのではなく、特定の化合物を出発点とする化合物生成手法の開発を目的とする。

2. 関連研究

Gomez-Bombarelli[5]らは VAE を用いて SMILES を生成することに成功した。しかし、生成された SMILES の多くは文法的に正しくないものであり、正しい SMILES を生成するために多くのステップを実行する必要があった。

Segler[6] らは RNN（Recurrent Neural Network）の一種である LSTM を利用することにより、高い割合で有効な SMILES を生成することに成功した。この手法は有効率が高かったものの、特定の物性の最適化という面で見ると効率が良いとは言えなかった。

特定の化合物を出発点とした化合物生成手法として、Zhou[7]らは深層 Q 学習を用いた分子グラフベースの分子生成手法を開発した。可能な行動を化学的に正しいもののみに制限することにより常に有効な化合物を生成することに成功したが、生成することが不可能な化合物も存在した。

また、複数の機械学習手法を組み合わせた手法も多く存在している。X.Yang[8]らは MCTS（モンテカルロ木探索）と RNN を組合わせた手法として ChemTS を開発した。ChemTS は VAE や RNN などの既存手法と比較して効率よく化合物を生成することに成功したが、特定の化合物を出発点とした化合物生成に対応していなかった。

† 東京工業大学情報理工学院 Tokyo Institute of Technology, School of Computing
‡ 東京工業大学物質・情報卓越教育院 Tokyo Tech Academy for Convergence of Materials and Informatics

3. 手法

本手法は MCTS と RNN を用いた、特定の化合物を出発点とした SMILES ベースの化合物生成モデルである。MCTS と RNN で部分的な SMILES を生成し、出発点の SMILES の一部と置き換えることにより新しい化合物を生成するという手法である。

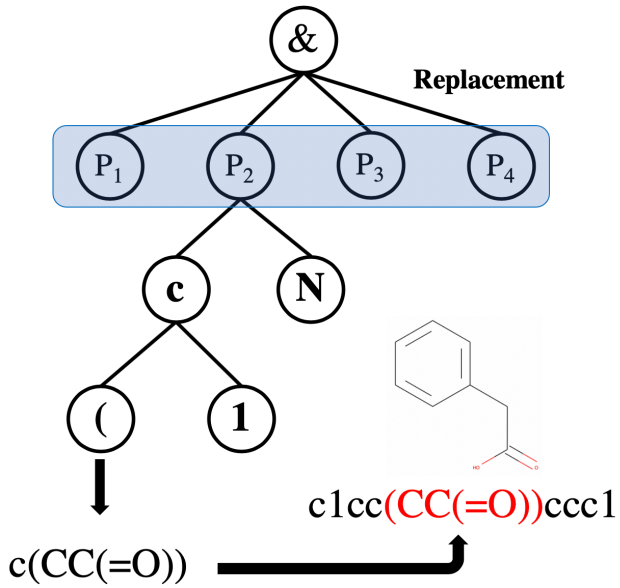

図 1 本手法による化合物生成の概略図

3.1 RNN

RNN は Neural Network の一種であり、層方向のみではなくデータ間方向にも情報が伝搬するモデルである。自然言語処理などの時系列データの特徴を捉えることに長けており、特に長期依存関係を得意とする LSTM など、様々な派生が提案されてきた。

本研究では分子の文字列表現である SMILES の特徴を捉えるために利用される。

3.2 MCTS (モンテカルロ木探索)

MCTS[9]とは強化学習アルゴリズムの一種であり、探索木として表現される。一つのノードは一つの状態に対応し、その価値と訪問回数を記憶する。探索は深さ優先で行われるため、探索されるノードに偏りが現れる。訪問するノードを適切に選択することにより、MCTS は効率の良い探索を実現することが可能である。具体的に、MCTS の探索は展開するノードの選択、ノードの展開、シミュレーション、

バックアップを 1 ステップとして、これを繰り返すことにより探索している。

本研究では、SMILES の一文字を 1 ノードに対応させている。これによって、ルートノードからリーフノードまでのパスが(部分的な)SMILES として表される。

モンテカルロ木探索は非終端ノードを評価するためにロールアウトを行う。ロールアウトとは評価したいノードから仮想的に終端ノードまで展開し、その終端ノードに対する評価を用いる方法である。本手法ではノードの展開に SMILES を学習した RNN を用いることにより精度の高い評価を行なっている。

3.3 部分 SMILES の置換

本手法では生成した部分 SMILES を元の SMILES の一部と置き換えることにより新しい化合物を生成する。この時置換される部分の選択は MCTS 含まれている。具体的には図 1 のようにルートノードの下に置換される部分 SMILES を表すノードが加えられ、望ましい化合物が生成されると期待される箇所が優先的に取り除かれる。

4. 学習データと実験内容

実験に使用される化合物データは ChEMBL から得られた 1,333,037 個の化合物である。その内 9 割が RNN の学習データとして用いられ残りが生成モデルの評価に用いられる。

実験の最適化対象は QED である。QED とは薬らしさを表す指標であり、0 から 1 の間で 1 に近いほど薬らしいことを意味する。実験は二つのモデルに対して行われる。一つは出発点となる化合物を固定する Single モデルであり、もう一つは出発点となる化合物が一定のステップ間隔で置き換えられる Mulit モデルである。テスト化合物は ChEMBL から得られた化合物 20 個(QED が 0.5 以下)であり、それぞれ 50,000 ステップ実行される。ただし、Multi モデルは 10,000 ステップ間隔で出発点となる化合物がそれまでに生成された化合物の中から最も QED が高いものに置き換えられる。

5. 実験結果

表 1　テストデータ化合物に対する 平均の QED

	テストデータ	Single	Multi
Top 1	0.3967	0.7789	0.9456
Top 2	—	0.7716	0.9449
Top 3	—	0.7659	0.9445

表 2　生成化合物全体の評価

	Validity	Uniqueness	Novelty
Single	0.1877	0.9811	1.0
Multi	0.2043	0.9906	1.0

本手法の最適化に対する評価が表 1 に示されている。Single モデルは上位の化合物の平均 QED が約 0.77 と上昇はしているものの十分ではないことがわかる。一方、Multi モデルは約 0.94 と十分な最適化が行われており、最適化という側面で見ると Multi モデルの方が優れていることがわかる。これは Single モデルが元の SMILES の一箇所のみが変化するのに対し、Multi モデルは元の SMILES が入れ替わるので結果的に複数箇所に変化が生じる SMILES を生成することが可能だからだと考えられる。一般的な生成モデルとしての評価は表 2 からわかる。Single、Multi ともに Validity は低いものの Uniqueness、Novelty は優れた結果が得られている。

図 2 に Multi モデルによって生成された化合物の構造を示している。左上の化合物は出発点となった化合物であり、多様な構造を持った優れた化合物が生成されていることがわかる。

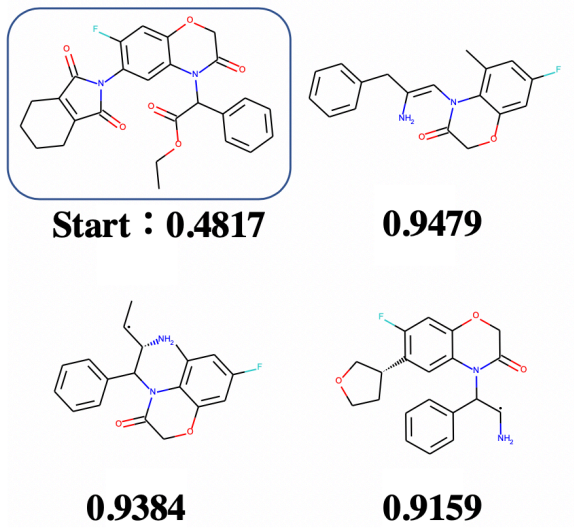

図 2　生成された化合物の構造と QED

6. 結論

本研究では特定の化合物を出発点とした誘導体生成手法を開発した。本手法は、モンテカルロ木探索と RNN を用いて部分的な SMILES を生成し、それを元の SMILES の一部と置き換えることによって新しい化合物を生成するという手法である。Single、Multi の二つのモデルを提案したが特に Multi モデルは Validity を除いて優れた性能を示し、最適化という観点からも十分な結果を示した。

参考文献

[1] Mullard,A. New drugs cost US $2.6 billion to develop, Nature Reviews Drug Discovery, Vol.13, No.877(2014)
[2] Keserü, G., Makara, G. The influence of lead discovery strategies on the properties of drug candidates. Nature Reviews Drug Discovery. Vol.8 (2009).
[3] Dobson, C. Chemical space and biology. Nature (2005)
[4] Sanchez-Lengeling,Aspuru-Guzik. Inverse molecular design using machine learning: Generative models for matter engineering. Science (2018)
[5] Gómez-Bombarelli, et al. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules.ACS Cent. Vol.4, No.2, (2018)
[6] Segler et al. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. ACS Cent. Vol.4, No.1, (2018)
[7] Zhou, Zhenpeng et al. Optimization of Molecules via Deep Reinforcement Learning. Scientific Reports. Vol.9, No.1, (2019)
[8] Yang, Xiufeng et al. ChemTS: an efficient python library for de novo molecular generation. Sience and Technology of Advanced Materials. Vol.18, No.1, (2017)
[9] Coulom, R. Efficient Selectivity and Backup Operators in Monte Carlo Tree Search. IEEE, Vol.4, No.1, (2012)