Exploration of molecular reaction networks behavior space with VAE-assisted Quality-Diversity

理学専攻・情報科学コース　1940660　山崎 瑛梨佳　(指導教員：オベル加藤 ナタナエル)

1　はじめに

分子計算に対する設計論として分子プログラミングの研究成果が多く発表されてきた [1][2]。分子計算とは生体分子が潜在的に持つ計算様式を分析し制御することを目的とした研究領域であり、主に DNA、RNA、タンパク質の化学反応に応用されている。この非常に複雑な計算機構を情報科学の観点から解釈しモデル化することで生体分子の化学反応の設計論を確立することを目的とした特定領域が分子プログラミングである。

しかし、生体における化学反応は精密なネットワークによって表現されるためプログラミングを行うことは非常に困難である。そこで本研究では、Variational Autoencoder(VAE)[3] と MAP-Elites[4] を適用することでその複雑性の課題に取り組んだ。VAE は深層生成モデルであり、エンコーダとデコーダの中間層である潜在空間におけるデータの次元削減が行われる。MAP-Elites は Quality-Diversity アルゴリズム (またはイルミネーションアルゴリズム) の一つで、新しいタイプの探索アルゴリズムである。近年では、科学や工学など様々な分野で探索アルゴリズムを使用している。探索アルゴリズムにより予測は困難だがパフォーマンスの高い解を見つけることができるため、分子の空間を探索することで新薬を発見したり、より優れたモデルのロボットの設計を探索したりということが可能である。中でもこの MAP-Elites は、従来のイルミネーションアルゴリズム (探索空間に分布する解のパフォーマンスを全体的に把握することを可能にする探索アルゴリズム) よりも実装が簡単で理解しやすい上、よりパフォーマンスの高い解を発見するのに役立つことが確認されている。

このような VAE の特性と MAP-Elites の機能を組み合わせることにより分子反応の探索を効果的に行うシステムを提案する。

2　モデル

時系列データに対して VAE で次元削減、その潜在空間に Quality-Diversity アルゴリズムを組み合わせ探索する実験を行った。この時系列データとは、PEN DNA toolbox[5] により分子反応ネットワークを生成し、それらを DNA Artificial Circuits Computer-Assisted Design (DACCAD)[6] に渡すことでシミュレーションを行い得られたものである (図 1)。これにより分子反応のアナログ表現と時系列データを変換している。緑色で表されるインヒビターは、特別な DNA シグナル鎖を介して行われる抑制モジュールで、入力には含まれないため、赤色ノードの数をネットワーク全体のノード数と考える。

このような時系列データに対して VAE を適用した。本研究で VAE は Keras を用いて実装している。用意したデータセットは、分子反応ネットワークのサイズが 1 から 5、長さ 500 の時系列とした。これは時系列が特徴的な挙動を示すのに十分な長さを考慮している。

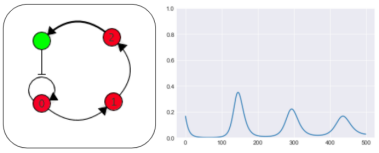

図 1: 分子反応ネットワークとそのシミュレーション結果の例

従って、これらの実験において 500 次元を d 次元に圧縮する。機能性と圧縮率を評価する実験により次元を決定する。そして時系列の特徴を反映したその潜在空間について Quality-Diversity アルゴリズムを運用する。ここで化学反応ネットワークを取得しそのシミュレーション結果の評価を経て探索結果を得る。これにより複雑なネットワーク表現に対するデータ探索を実現した。

3　実験結果

3.1　次元削減

VAE による次元削減を効果的にするため、潜在空間が 2 次元から 10 次元の場合について VAE の性能を評価する実験を二つの観点で行った。VAE によるデータの復元がうまく機能しているか、さらに VAE の潜在空間における分布が妥当に行われているか、である。まず、用意したデータセットで訓練した VAE に、3 種類のサンプルデータをエンコードし、生成された (デコードされた) データの再現性を評価した (図 2)。青色が入力したサンプルデータ、オレンジ色が生成されたデータを表す。実験に用いたサンプルデータのうち、単調増加関数と振動子についてデコードした結果を抜粋している。訓練データは、ネットワークのサイズを固定し (1 から 5) ランダムに反応を発生させ作成している。また、そこにネットワーク数 3 で振動の特徴を持つデータも加えることで多様性のあるデータセットを用意している。2 次元のように次元が低いと、特徴

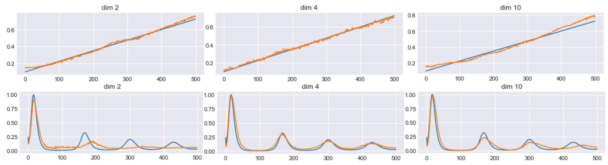

図 2: VAE のデコード結果

的な挙動 (例えば振動子のピーク) さえ再現されていない。一方で、単純に次元を増しても再現がうまく機能しない結果が確認された。4 次元に比べ 10 次元の結果は再現率が低い。次に、潜在空間への圧縮が効果的かを定量的な分析により評価した。一般的に用いられる次元削減手法である PCA と比較実験を行った。次元削減された VAE の潜在空間と PCA の特徴空間に各々 K-means 法によりクラスタリングすることでデータを

5 つに分類し、それに含まれる時系列の特徴を分析した。テストデータにおいて単調増加・減少関数、定数、振動子、sigmoid 関数の振る舞いを見せるデータについて分類を計上した (表 1・図 3)。これらの特徴は分子反応ネットワークにおける実験結果として一般的によく観測されるため選択している。PCA は次元に依らず安定した結果で、特徴量 3 次元に圧縮した場合で最も高い 62.4%の正答率を示した。一方 VAE は次元により性能にかなり差が生じ、4 次元の結果は 65.6%で PCA と VAE を合わせた中で最も高い正答率を示した。これらの分類を可視化したのが図 3 である。色は時系列の種類ごとに分類された割合を表している。VAE は振動子と sigmoid 関数の特徴を持つ時系列を著しく正確に分類している。増加・減少関数と定数の時系列については近いラベリングが行われている。ここで注目したいのは、VAE の結果には色の薄い (100 かそれに近い) セルが多くなっているということだ。つまり、VAE では分類がより明確に成されている。

表 1: 分類の正答率 (%)

次元	3	4	5	6	7	8
PCA	**62.4**	55.2	56.0	58.4	57.6	57.6
VAE	44.0	**65.6**	55.2	55.2	53.6	**62.4**

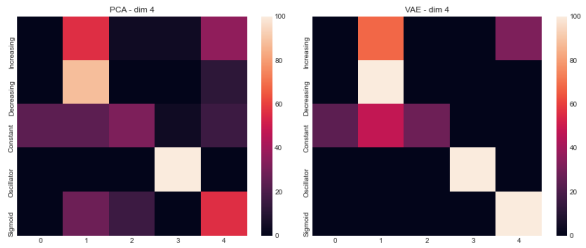

図 3: 時系列における分類の可視化 (4 次元)

そこで、分類がどのように機能しているか詳細を示したのが図 4 である。時系列における分類を色で表現している。左の PCA による結果ではラベリングが値域に基づいている傾向が見られるが、VAE では定数値をとる時系列についてはその値による分類が行われている。以上の実験から、VAE の復元機能と潜在空間におけるクラスタリングの正当性の評価から、VAE の潜在空間を以下 4 次元と定義する。

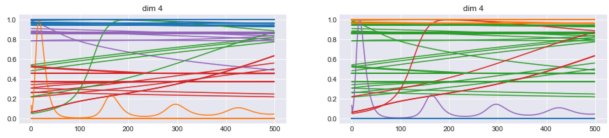

図 4: 時系列表現におけるラベリング (4 次元)

3.2 データ探索

Qdpy[7] により実現した MAP-Elites スタイルのアルゴリズムをこの VAE に適用した。このアルゴリズムは、古典的な進化アルゴリズムが母集団を使用するのと同様にコンテナ (グリッド) を使用し、探索した個体の突然変異から得られた新しい解で反復的にグリッドを埋めていく。これに目的関数を設定し、VAE の潜在空間で探索を行った (図 5)。広範囲に渡り分布する最適解を探索できていることが確認できる。また、ビンの更新回数も計算量を鑑みても妥当である。なお、このグリッドを表示する過程で DACCAD によるシミュレーションを行っているため、色付けのされていないセルが存在するのは PEN DNA toolbox で反応ネットワークの表現が存在しない場合があるためである。

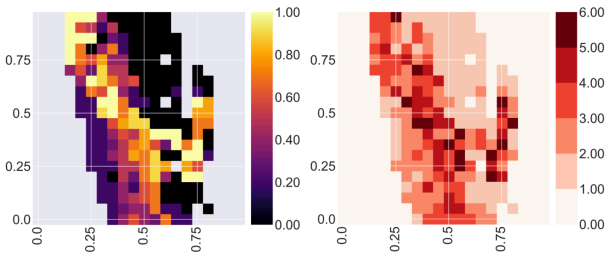

図 5: 潜在空間における探索例

4 まとめ

本論文では、分子プログラミングの領域におけるデータ探索を行った。化学反応の時系列データを対象に効果的な次元削減を行い Quality-Diversity アルゴリズムを適用することで、高次元な数学的対象におけるデータの探索を可能にした。時系列データの特徴を持つ潜在空間について探索を進めると同時に、化学反応ネットワークにおけるシミュレーション結果を評価しているため、化学反応の実現性に基づいた探索が実現される。このシステムが実用的に機能していることが確認された一方で、システムの部分的な向上の余地を残している。また、この研究で用いたデータや評価関数は試験的なものであるが、化学的実験によるデータに適用や有効なカスタマイズといった大域的な拡張が期待される。

参考文献

[1] Lulu Qian and Erik Winfree. Scaling up digital circuit computation with dna strand displacement cascades. *Science*, Vol. 332, No. 6034, pp. 1196–1201, 2011.

[2] Adrien Padirac, Teruo Fujii, and Yannick Rondelez. Bottom-up construction of in vitro switchable memories. *Proceedings of the National Academy of Sciences*, Vol. 109, No. 47, pp. E3212–E3220, 2012.

[3] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2014.

[4] Jean-Baptiste Mouret and Jeff Clune. Illuminating search spaces by mapping elites, 2015.

[5] Alexandre Baccouche, Kevin Montagne, Adrien Padirac, Teruo Fujii, and Yannick Rondelez. Dynamic dna-toolbox reaction circuits: A walkthrough. *Methods*, Vol. 67, No. 2, pp. 234 – 249, 2014. Nucleic Acids Nanotechnology.

[6] Nathanaël Aubert, Clément Mosca, Teruo Fujii, Masami Hagiya, and Yannick Rondelez. Computer-assisted design for scaling up systems based on dna reaction networks. *Journal of The Royal Society Interface*, Vol. 11, No. 93, p. 20131167, 2014.

[7] L Cazenille. Qdpy: A python framework for quality-diversity, 2018.