

Convolutional Neural Network によるヒット化合物判定

杉山佳奈子^{†1}, 浅谷学嗣^{†1}, 田川聖一^{†1}, 新岡宏彦^{†1},
 齊藤隆太^{†2}, 西村裕一^{†2}, 松本健太郎^{†2}, 大野研^{†2}, 三宅淳^{†1}

概要 : 創薬におけるフェノタイプスクリーニングの手法として Deep Learning を適用した. ある化合物を細胞に添加した際に細胞の表現型が変化したかどうかを Deep Learning を用いて判定することを試みた. 複数の化合物において目視では判別が困難な表現型の変化を自動的に検出することができた. 高い精度で判別できた化合物とそうでないものがあることから, この手法を用いることで, 細胞に何らかの影響を与える化合物 (ヒット化合物) を検出することができるのではないかと考える.

キーワード : Deep Learning, Convolutional Neural Network, フェノタイプスクリーニング

1. 背景・目的

医薬品となる化合物を発見するために行われるスクリーニングの一つに, フェノタイプスクリーニングがある. これは細胞や臓器の表現型の変化によって化合物の細胞への影響を評価する手法であり, 特定の分子のみを標的としたスクリーニング手法 (target-based screening) に代わり近年注目されている[1]. フェノタイプスクリーニングをはじめ, 創薬におけるスクリーニングでは膨大な情報を扱うため, 機械学習を用いた研究が数多くなされている[2][3]. また, 近年では創薬の研究に機械学習手法の一種である” Deep Learning ”を用いた事例もある[4]. Deep Learning はヒトの脳を模して造られた学習機構である. 従来の学習手法が手動で特徴量を設定するのに対し, Deep Learning では学習過程で特徴量を自動的に抽出することができることが特色であり, 画像認識[5]やチャットによる適切な会話の実現[6]など, 幅広い分野で成果が挙げられている. 本研究では細胞のフェノタイプスクリーニングを目的とし, Deep Learning の中でも画像解析に適した学習手法である Convolutional Neural Network (CNN) を用いて細胞の表現型の識別を行った. この CNN は, 画像認識の競技会である ILSVRC において, 2012 年以降の大会で優秀な成績を収めている[7]. 化合物が添加された際の細胞の表現型の特徴を Deep Learning を用いて捉えることで, 化合物が生体に与える影響を予測することができるようになり, 創薬研究に大きな貢献をすることができると思われる.

2. 方法

取得した細胞画像から機械学習に適した形式に整えた入力画像を作成し, 細胞に対する化合物添加の有無について CNN による学習を行った. 用いた細胞はヒトの骨肉腫細胞

表 1 本研究で用いた化合物

系列	化合物名
1	(-)-Blebbistatin
2	(-)-Epigallocatechin gallate
3	(-)-MK 801 Maleate
4	(+)-Bicuculline
5	(+)-Usnicin (D-Usnic acid)
6	(+,-)-Octopamine HCl
7	(R)-Nepicastat HCl
8	(S)-(+)-Flurbiprofen
9	(S)-crizotinib
10	(6-) ε -Aminocaproic acid
11	10058-F4
12	10-DAB (10-Deacetylbaecatin)
13	1-Azakenpaullone
14	1-Hexadecanol
15	20-Hydroxyecdysone
16	2-Thiouracil
17	3-deazaneplanocin A (DZNeP) HCl
18	3-Indolebutyric acid (IBA)
19	3-Methyladenine (3-MA)
20	4μ8C

由来である U2OS である. 用いた画像は 4 種類の蛍光画像で構成されており, それぞれ異なる細胞内小器官が撮影される (ch1 : 核, ch2 : ゴルジ体/細胞膜, ch3 : ER/核小体, ch4 : ミトコンドリア). 添加した化合物は 20 種類であり, 名称とグラフで用いる系列番号を表 1 に示す. 各化合物を添加したときの画像の学習結果を比較することで, 化合物ごとの細胞の表現型に与える影響を評価した.

2.1 入力画像の作成

元画像の ch1 である核画像において楕円検出を行い, 細胞中心を検出する. 元画像のすべてのチャンネルで輝度を正規化し, 細胞核を中心とした画像を切り出す (150×150px). 切り出した画像のうち 100 枚をテスト画像として用い (化合物添加有り 50, 添加無し 50), 残りを訓練画像として用いた.

^{†1} 大阪大学
 Osaka University
^{†2} 田辺三菱製薬株式会社
 Mitsubishi Tanabe Pharma Corporation

2.2 CNN の構成

用いた CNN の構成を図 1 に示す。入力は 150×150 の 4 チャンネル画像であり、0~1 の出力値に対して教師データは化合物添加有りを"1", 添加無しを"0"とした。

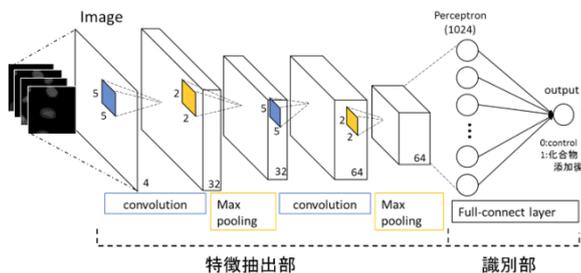


図 1 CNN の構成

3. 結果

20 種類の化合物を添加した細胞の画像についてそれぞれ学習を行なったところ、正答率 (テスト画像を入力した際の出力を四捨五入した際、教師データと一致した割合) は、添加した化合物の種類によって異なり、52%から 94%の値をとった。化合物ごとの正答率を表したグラフを図 2 に示す。また、20 種類のうち正答率の高い化合物と低い化合物のテスト画像 (左から順に ch1, ch2, ch3, ch4) を図 3 に示す。上から化合物添加無し, 正答率の高い化合物 3 種, 正答率の低い化合物 3 種の細胞画像となっている。正答率が低かった化合物の画像と同様に、正答率が高かった化合物の画像も目視での判別は難しいことがわかる。

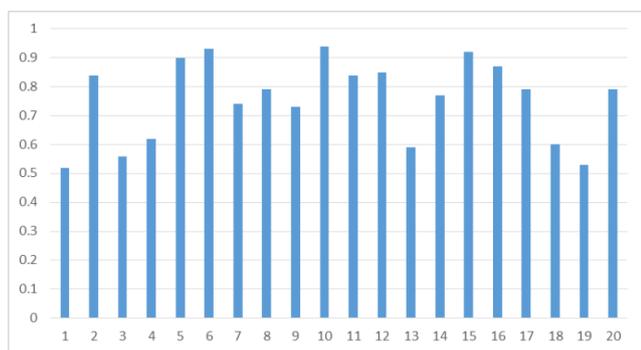


図 2 各化合物における正答率

4. 結論

目視での判別が難しい細胞画像を、複数のデータセットにおいて高い正答率で判別することができたため、これらはヒット化合物であると考えられる。一方、正答率が低かったデータセットについては細胞に影響を与えない化合物であったと考えられる。このように、添加した化合物の種類によって CNN による学習結果が大きく異なったことから、細胞の表現型の評価に CNN が有用であることが示唆された。

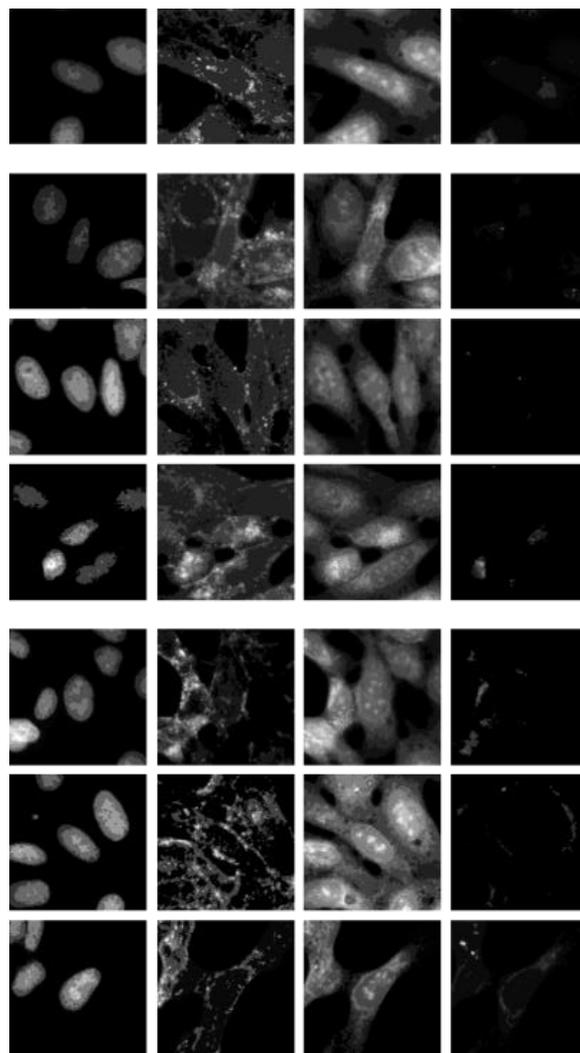


図 3 各化合物における添加後の細胞画像

参考文献

- [1] 市川聡, et al. "SDS-PAGE 解析 細胞周期 フェノタイプスクリーニング 生体直交性有機化学." 有機合成化学協会誌 74.5, pp.541-541 (2016).
- [2] Feng, Yan, et al. "Multi-parameter phenotypic profiling: using cellular effects to characterize small-molecule compounds." Nature Reviews Drug Discovery 8.7, pp.567-578 (2009).
- [3] Wawer, Mathias J., et al. "Toward performance-diverse small-molecule libraries for cell-based phenotypic screening using multiplexed high-dimensional profiling." Proceedings of the National Academy of Sciences 111.30, pp.10911-10916 (2014).
- [4] 浜中雅俊, et al. "ディープラーニングに基づく化合物とタンパク質の相互作用予測." ケモインフォマティクス討論会予稿集 2015.0, pp.46-49 (2015).
- [5] Vinyals, Oriol, and Quoc Le. "A neural conversational model." arXiv preprint arXiv:1506.05869 (2015).
- [6] Le, Quoc V. "Building high-level features using large scale unsupervised learning." IEEE international conference on acoustics, speech and signal processing. 2013.
- [7] Russakovsky, Olga, et al. "Imagenet large scale visual recognition challenge." International Journal of Computer Vision 115.3, pp.211-252 (2015).