Vol.2017-BIO-51 No.3
2017/9/26

ニューラルネットワークを用いた Gタンパク質共役受容体のクラスタリング

遠藤麗花，　長田直樹，　遠藤俊徳
北海道大学大学院　情報科学研究科　生命人間情報科学専攻　情報生物学研究室

研究背景

G タンパク質共役受容体（GPCR）は，細胞内情報伝達系最上流に位置し，新薬開発の標的分子の一つである．リガンドおよび機能が未知の GPCR を Orphan GPCR と呼ぶ．Orphan GPCR およびリガンドの解明は，新薬開発への貢献につながると期待される．我々はこれまで，機械学習の応用による嗅覚受容体および Class A GPCR のクラスタリングによる GPCR の機能推定を行ってきた[1][2]．Orphan GPCR に対し教師なし学習のひとつである自己組織化マップを適用し，その結果受容体のクラスタリングに成功した．

問題点

GPCR の特徴情報は次元数が非常に大きいものになり得るため，無選別に機械学習法に適用すると，分類精度が悪化すると予想される．

手段

畳み込みニューラルネットワークは通常のニューラルネットワークに畳み込み層，プーリング層と呼ばれる層を付加したもので，入力データの本質的な情報を抽出し，より精度の高い分類を実現する．この手法を GPCRdb[3] のデータに適用することでより精度の高いリガンド推定が可能になると期待される．

目的

本研究では，画像認識で利用される畳み込みニューラルネットワークを用いた GPCR の分類を行う．

方法

リガンド既知 GPCR データの抽出

本研究では，ニューラルネットワークによる分類を行うにあたり，GPCRdb から得られた 99 個の ヒト GPCR を考える（表1）．このデータには PDB ファイルから抽出した最大4165件の要素が含まれている．各受容体の遺伝子はアミノ酸数が異なるため，各データの要素数はすべて統一されていないことに注意する．PDB ファイルから抽出したデータは α 炭素原子の直交 X 座標，Y 座標，Z 座標，アミノ酸情報から Kyte-Doolittle[4]によって数値化した疎水性インデックスおよび分配係数を含めた．

教師データの作成

本研究では，Orphan GPCR のクラスタリングを教師あり学習手法を用いて実現するため，使用する教師データの出力を0，1とした．なお，Class A である GPCR を1，Class A 以外の GPCR を0としてラベリングした．リガンド既知の GPCR の分類を行うにあたり，教師データを作成した．

表1　学習に用いた GPCR の内訳

	GPCRの個数
Class A	87
Class B	3
Class C	4
Class F	5
合計	99

ニューラルネットワークの学習

本研究では，精度の高い分類を実現するため4層の畳み込みニューラルネットワーク（畳み込み層→プーリング層→全結合層1→全結合層2）を考える（図1）．訓練データとしてリガンド既知 GPCR データを用いて学習を行った．なお，多層ニューラルネットワークの訓練精度を比較するため，全結合層1および全結合層2で構成される2層ニューラルネットワーク（図2）を用いたリガンド既知 GPCR の分類も実施した．

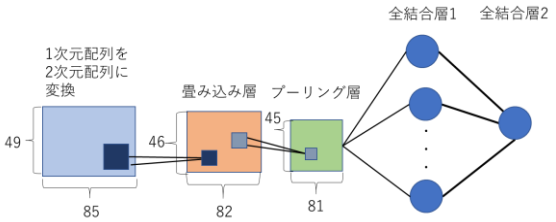

図1　4層畳み込みニューラルネットワークの概略

ⓒ 2017 Information Processing Society of Japan

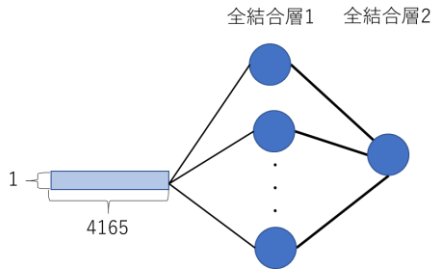

図2　2層ニューラルネットワークの概略

学習に用いた訓練データの各入力は 4165 要素の1次元配列であり出力は1次元である．なお，畳み込み層に入力するため1次元配列の入力データを49×85の2次元配列にリサイズした．

また，99個のGPCRの中から無作為に抽出した89個のGPCRを学習に用い，残りの10個のGPCRの分類を行う工程を10回行った．2層および4層ニューラルネットワークの訓練制度の検証を行った．4層畳み込みニューラルネットワークにおける学習に用いたパラメータを表2に示す．

表2　4層畳み込みニューラルネットワークパラメータ

4 層ニューラルネットワーク	
層数	4
重み（初期値）	標準偏差 1.0×10^{-4}，平均 0 の正規分布(畳み込み層，全結合層)
バイアス（初期値）	1.0×10^{-4}(畳み込み層，全結合層)
フィルター数	1
サイズ	4×4（畳み込み層），2×2（プーリング層）
ストライド	2（畳み込み層，プーリング層）
入力チャネル数	1
出力チャネル数	1
ゼロパディング	なし（畳み込み層，プーリング層）
プーリング方法	Max Pooling
ユニット数	10(全結合層1)，1(全結合層2)
学習係数	1.0×10^{-8}
活性化関数	恒等写像（畳み込み層，全結合層1），シグモイド関数（全結合層2）
optimizer	勾配降下法
誤差関数	二乗誤差
教師データ	2 値(0，1)

結果

リガンド既知 GPCR の分類結果

2層ニューラルネットワークおよび4層畳み込みニューラルネットワークが精度よく学習されていることを確認するため，学習に使用していないリガンド既知の GPCR データ10個を分類し，クロスバリデーションで評価した．2層ニューラルネットワークを用いてリガンド既知の GPCR の分類を10回行ったところ，クロスバリデーションによる正解率（accuracy)の平均は0.75で false positive rate の平均は0.10だった．一方，4層の畳み込みニューラルネットワークを用いた場合の平均 accuracy は0.91で false positive rate の平均は0.09だった．

考察

リガンド既知の GPCR の分類では取得したテストデータにおいて Class A GPCR は正しく分類されたが過学習が発生している。しかし、2層のニューラルネットワークにおいて4層とほぼ同じ false positive rate が得られたため，全結合層は過学習を起こしてないことが考えられる．

また，テストデータの中で Class A 以外の GPCR の数がクロスバリデーションの accuracy に影響を与えることが考えられるが、2層と比較したときの4層における accuracy が0.16向上して false positive rate が0.01減少したため，ネットワークを多層にして分類することに意味があるといえる．

まとめ

本研究では，4層の畳み込みニューラルネットワークを用いて GPGR の分類を行った．リガンドが既知の GPCR の分類では2層のニューラルネットワークを用いた予測結果は正解率が0.75であったが，4層の畳み込みネットワークでは0.91と正解率が向上し、false positive rate は 0.01下がった．

しかし，Class A 以外のデータが少ないためさらなる入力データを用いた検証が必要である．

参考文献

[1] 小坂,北海道大学大学院修士論文(2008)
[2] 吉田,北海道大学大学院修士論文(2015)
[3] http://gpcrdb.org/
[4] Kyte,J;Doolitle,R.F, J.Mol.Biol.,157(1),105-32(1982)

ⓒ 2017 Information Processing Society of Japan

正誤表

下記の通り，誤記がありましたので訂正いたします.

正誤箇所	誤	正
2 ページ-表 2 ストライド	2（畳み込み層，プーリング層）	1（畳み込み層，プーリング層）