4ZC-02

糖鎖情報を組み込んだニューラルネットワークによる薬剤とタンパク質の相互作用予測

新川　栄二†　　永塚　光一††　　村田　祐樹††　　小野　多美子†††
細田　正恵†††　　木下　聖子†††　　渥美　雅保††
†創価大学理工学部情報システム工学科　　††創価大学大学院理工学研究科
†††創価大学糖鎖生命システム融合研究所

1. はじめに

近年，医薬データ等の増加に伴い創薬分野における機械学習を用いた研究が盛んに行われている．特に標的物質と薬剤の相互作用を予測することは既存薬剤の新規効能を発見するドラッグリポジショニングにのみならず，その相互作用部位を解析することで新薬開発においても有用である．本研究では，第三の生命鎖として注目されている糖鎖に着目し，薬剤と糖タンパク質の相互作用予測において，(1)薬剤，アミノ酸配列情報，(2)薬剤，糖鎖情報，(3)薬剤，アミノ酸配列情報，糖鎖情報の 3 つの組み合わせを入力とするニューラルネットワークモデルの精度を比較する．そして実験結果から，相互作用予測性能における糖鎖情報の有用性を示す．

2. 関連研究

機械学習による創薬の工程は，薬剤の標的となるタンパク質等の物質を探索する工程と探索した物質に対する標的薬を選択・開発する工程に分けられる．後者の研究では薬剤と対象タンパク質の相互作用を 2 つの深層学習モデルを組み合わせて予測する手法が提案されている[1]．この手法は薬剤とタンパク質の表現としてそれぞれ SMILES という線形表記とアミノ酸 n-gram 配列を得た後に，グラフニューラルネットワークと畳み込みニューラルネットワークで特徴ベクトルを抽出し，それらを結合した上で相互作用を予測する．本研究では，このモデルを上記(1)薬剤，アミノ酸配列情報を入力とする「アミノ酸配列モデル」として比較実験に用いる．

3. 提案手法

先行研究の「アミノ酸配列モデル」をベースに開発した上記(2)薬剤，糖鎖情報を入力とする「糖鎖モデル」，(3)薬剤，アミノ酸配列情報，糖鎖情報を入力とする「複合モデル」について述べる．

3.1. 糖鎖頻度ベクトル生成

糖鎖モデルと複合モデルでは糖鎖情報として糖鎖頻度ベクトルを用いる．糖鎖頻度ベクトルは大きく分けて，該当するタンパク質に付随する糖鎖群における単糖を数えた単糖単位の頻度ベクトルと単糖構成を数えた糖鎖単位の頻度ベクトルの 2 つを考えることができる．また，それぞれ分子情報を考慮するかどうかの 2 種類が考えられるため，計 4 種類の糖鎖頻度ベクトルを用いる．

3.2. 提案モデル

3.2.1. 糖鎖モデル

糖鎖モデルを図 1 に示す．入力は SMILES と糖鎖頻度ベクトルである．糖鎖エンコーダーは 3 層のフィードフォワードニューラルネットワークで構成されており，抽出した特徴ベクトルをクエリとして薬剤のサブグラフベクトルに対するアテンションをとる．その結果出来上がった糖鎖特徴ベクトル，薬剤特徴ベクトルを結合し，線形変換をすることにより，相互作用ありなしの二値分類を行う．

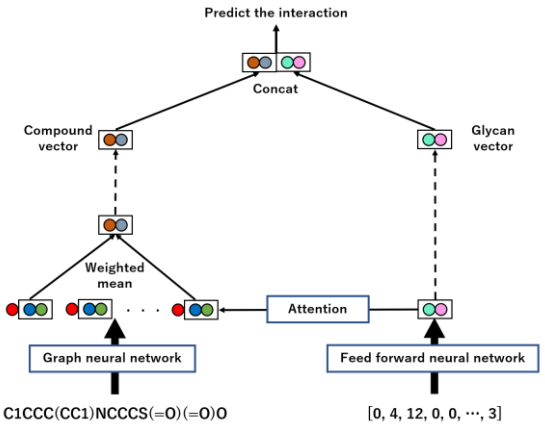

図 1　糖鎖モデル

3.2.2. 複合モデル

複合モデルでは，アミノ酸配列情報と糖鎖情報の組み合わせ方により 3 つのモデルを構築する．

3.2.2.1. 相互アテンションモデル

相互アテンションモデルを図 2 に示す．入力は SMILES とアミノ酸配列ベクトルと糖鎖頻度ベクトルの 3 つである．本モデルは，糖鎖が薬剤とタンパク質を仲介するプロセスをモデル化している．エンコーディングした糖鎖特徴ベクトルをクエリとして薬剤のサブグラフベクトルとのアテンションを取った後に，薬剤特徴ベクトルをクエリとしてアミノ酸 n-gram 畳み込みベクトルとのアテンションをとる．その結果出来上がったアミノ酸配列特徴

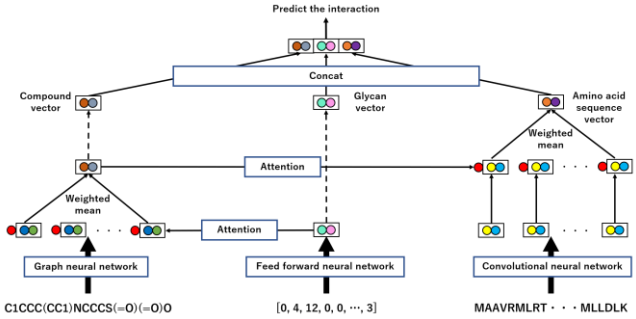

図 2　相互アテンションモデル

Compound-protein Interaction Prediction based on Neural Networks incorporating Glycan Information
Eiji Shinkawa†, Koichi Nagatsuka††, Yuki Murata††, Tamiko Ono†††, Masae Hosoda†††, Kiyoko Kinoshita†††, Masayasu Atsumi††
†Faculty of Science and Engineering, Department of Information Systems Engineering, Soka University
††Graduate School of Science and Engineering, Soka University
†††Glycan & Life System Integration Center, Soka University

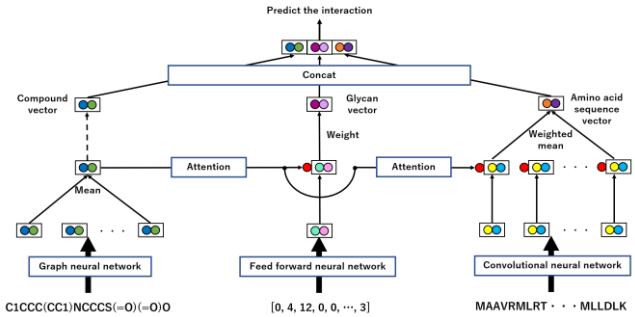

図 3　比較用モデル

ベクトル，糖鎖特徴ベクトル，薬剤特徴ベクトルを結合し，線形変換をすることにより，相互作用ありなしの二値分類を行う．

3.2.2.2. 比較用モデル

比較用モデルを図 3 に示す．入力は SMILES とアミノ酸配列と糖鎖頻度ベクトルの 3 つである．本モデルは 3.2.2.1 の糖鎖が薬剤と糖タンパク質を仲介するプロセスをモデル化した相互アテンションモデルとのアテンションのかけ方の比較用に用意したモデルである．エンコーディングした薬剤特徴ベクトルをクエリとして糖鎖特徴ベクトルと，アミノ酸 n-gram 畳み込みベクトルの 2 つとのアテンションをとる．その結果出来上がったアミノ酸配列特徴ベクトル，糖鎖特徴ベクトル，薬剤特徴ベクトルを結合し，線形変換をすることにより，相互作用ありなしの二値分類を行う．

3.2.2.3. アンサンブルモデル

アミノ酸配列モデルと糖鎖モデルで出力した相互作用のスコアをある重みで加重和したスコアをもとに相互作用ありなしの二値分類を行う．

4. データセット

4.1. 相互作用データセット

薬剤-糖タンパク質間に相互作用があるペアサンプルを，タンパク質との相互作用情報などを含む薬物に関する臨床レベルのデータを集めたデータベース DrugBank[2]や信頼できるネガティブサンプルを作成した研究[3]から集めた．相互作用データは相互作用ありデータが 1174，相互作用なしデータが 2348 の 1:2 の割合で用意した．

4.2. 糖鎖情報データセット

WURCS[4]という単糖構成などの情報を含む糖鎖情報の線形表記データを糖タンパク質などの複合糖質に関する情報を含んだ各種のリソースを整理した統合データベース GlyGen[5]から抽出した．

5. 実験結果

本研究では，アミノ酸配列モデル，糖鎖モデル，複合モデルの 3 つを同じデータセットを用いて比較実験する．10-分割交差検定で行った実験結果の内，アミノ酸配列モデル，糖鎖モデル，複合モデルの中でそれぞれの最高値を出したモデルの比較を表 1 に，さらに全モデルの中で最高値を出した相互アテンションモデルに関して，4 種類の糖鎖頻度ベクトルに対する予測性能比較を表 2 に示す．これらの実験結果が示すように，糖鎖単位かつ分子情報ありとして表現される糖鎖頻度ベクトルとアミノ酸配列に基づく相互アテンションモデルの予測性能が全モデルの中で 89.1%と最も高かったことが分かった．また，アミノ酸配列モデルよりも糖鎖モデルの方がより高い予測性能を示しており，中でも糖鎖頻度ベクトルが糖鎖単位かつ分子情報なしとして表現されるモデルが 86.7%と最も高かった．

表 1　3 モデルの最高値比較

モデル名	単位	分子情報	糖鎖比	AUC	precision	recall
アミノ酸配列モデル	-	-	-	0.826	0.653	0.745
糖鎖モデル	糖鎖	なし	-	0.867	0.725	0.756
相互アテンションモデル	糖鎖	あり	-	0.891	0.733	0.773
比較用モデル	糖鎖	なし	-	0.880	0.722	0.779
アンサンブルモデル	糖鎖	なし	0.9	0.876	0.724	0.755

表 2　相互アテンションモデルの糖鎖頻度ベクトル比較

単位	分子情報	AUC	precision	recall
糖鎖	なし	0.854	0.674	0.765
糖鎖	あり	0.891	0.733	0.773
単糖	なし	0.849	0.712	0.728
単糖	あり	0.857	0.691	0.738

6. 考察

本研究で用いたデータセットとモデルでは糖鎖情報を用いることでより高い相互作用予測精度が得られることが分かった．また，アミノ酸配列モデルや糖鎖モデルと比較して，両者を組み合わせた複合モデルでより高い予測精度が得られた．これらのことから相互作用予測において糖鎖がアミノ酸配列より有効である可能性，また糖鎖とアミノ酸配列でそれぞれ違う相互作用性を持っている可能性がある．さらに複合モデルでは相互アテンションモデルが最も予測性能が高かったことから，糖鎖が薬剤とタンパク質を仲介するプロセスをモデル化することで予測性能を向上させることができることが示唆された．

7. まとめ

本研究では，(1)薬剤，アミノ酸配列情報を用いるアミノ酸配列モデル，(2)薬剤，糖鎖情報を用いる糖鎖モデル，(3)薬剤，アミノ酸配列情報，糖鎖情報を用いる複合モデルを比較実験し，糖鎖情報の有用性を示した．さらに糖鎖が薬剤とタンパク質を仲介するプロセスをモデル化することにより，より高い予測性能が得られることが分かった．今後の課題として，相互アテンションモデルの改良が挙げられる．

参考文献

[1] Tsubaki, M., *et al.*, Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences, Bioinfomatics, Vol.35, No.2, pp.309-318, 2019.
[2] Wishart, D. S., *et al.*, DrugBank 5.0: a major update to the DrugBank database for 2018, Nucleic Acids Res, Vol.46, No.D1, D1074-D1082, 2018.
[3] Liu, H., *et al.*, Improving compound-protein interaction prediction by building up highly credible negative samples, Bioinformatics, Vol.31, No.12, pp.i221-i229, 2015.
[4] Tanaka, K., *et al.*, WURCS: The Web3 Unique Representation of Carbohydrate Structures. Journal of Chemical Information and Modeling, Vol.54, No.6, pp.1558-1566, 2014.
[5] York, W.S., *et al.*, GlyGen: computational and informatics resources for glycoscience, Glycobiology, Vol.30, No.2, pp.72-73, 2020.