5E-01

ニューラルネットワークによる糖タンパク質からの遺伝子変異量予測

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

1.　はじめに

遺伝子変異により正常とは異なるアミノ酸配列のタンパク質が作られることから，タンパク質と遺伝子変異の関係の解析は分子標的薬の開発において重要である．こうしたタンパク質の異常と潜在的な関連性があることが分かっている物質として，近年糖鎖が注目を集めている[1]．しかしながら，糖鎖を活用して，タンパク質と遺伝子変異の関係をデータ駆動型のアプローチにより予測する研究はまだ十分に行われていない．そこで本研究では，糖タンパク質に注目し，癌細胞を対象にして，糖鎖とアミノ酸配列からの遺伝子変異量の予測をニューラルネットワークにより行うモデルを提案し，遺伝子変異量の予測における糖鎖との関係を明らかにする．実験により，提案モデルの予測と実際の遺伝子変異量との間の順位相関において高い正の相関が得られ，糖鎖情報を組み合わせることにより予測精度が向上することを確認した．

2.　提案手法

2.1.　遺伝子変異量予測タスク

本研究における遺伝子変異量予測タスクとは，糖タンパク質のアミノ酸配列と糖鎖情報を入力し，遺伝子変異量を出力するタスクである．タンパク質情報の入力では，先行研究[2]と同様にアミノ酸配列を n-gram に分割し，n-gram アミノ酸ワード配列としてモデルに与える．糖鎖情報の入力では，糖タンパク質に対応する WURCS 表記[3]の糖鎖情報を元に，糖鎖と単糖の 2 つの粒度においてそれぞれの頻度をカウントし，そこから得られる頻度ベクトルをモデルへと入力する．

一方，予測する遺伝子変異量は，糖タンパク質の遺伝子変異数(Mutation Count)に対し，その糖タンパク質の発現量(Expression Level)を掛けた値として定義される．実験では，学習の安定性のために 0 から 1 の範囲で正規化処理を施した遺伝子変異量の値を利用する．

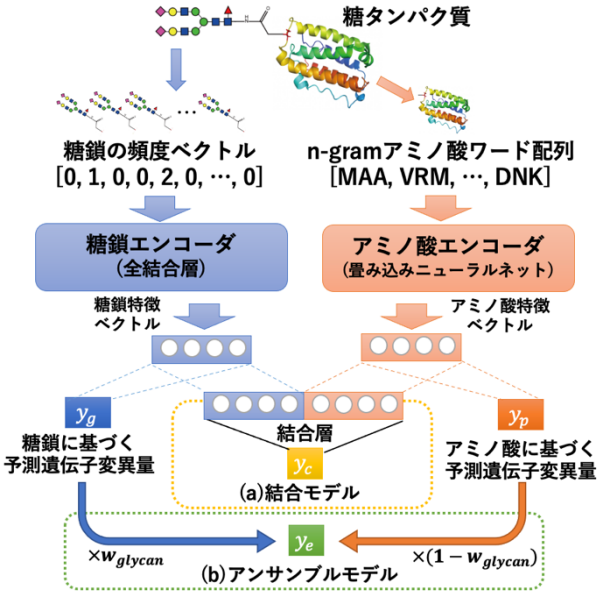

図 1 糖鎖とアミノ酸に基づく提案モデル

2.2.　提案モデル

図 1 に提案モデルのアーキテクチャを示す．提案モデルはアミノ酸エンコーダと糖鎖エンコーダの 2 つのコンポーネントから構成される．アミノ酸エンコーダは多層の畳み込みニューラルネットワークで構成されており，n-gram で分割されたアミノ酸の埋め込みベクトルを受け取り，アミノ酸特徴ベクトルを出力する．糖鎖エンコーダは，多層の全結合層からなり，糖鎖の頻度ベクトルを受け取り，糖鎖特徴ベクトルを出力する．図 1(a)の結合モデルはこれらの特徴ベクトルを結合し，線形変換により予測遺伝子変異量を求める．一方，図 1(b)のアンサンブルモデルは，2 つエンコーダの出力を線形変換して求めた予測遺伝子変異量に対し重み付けを行い，最終的な予測遺伝子変異量を決定する．

3.　データセット

3.1.　The Cancer Genome Atlas

The Cancer Genome Atlas(TCGA)は，癌研究に広く利用されているデータベースであり，33 種類の癌種について，ゲノム変異や遺伝子発現変動などの情報を公開している．実験においては，糖タンパク質に対して，遺伝子変異数(Mutation Count)を抽出し，データセットを構築した．

Mutation Count Prediction from Glycoprotein based on Neural Networks
Koichi Nagatsuka[†], Yuki Murata[†], Eiji Shinkawa[††], Tamiko Ono[†††], Masae Hosoda[†††], Kiyoko Kinoshita[†††], Masayasu Atsumi[†]
[†]Graduate School of Science and Engineering, Soka University
[††]Faculty of Science and Engineering, Department of Information Systems Engineering, Soka University
[†††]Glycan & Life System Integration Center, Soka University

表 1 各モデルにおける順位相関係数(太字は各試行及び平均における最高値)

Model	糖鎖の重み	アミノ酸の重み	Trial-1	Trial-2	Trial-3	Trial-4	Trial-5	Trial-6	Trial-7	Trial-8	Trial-9	Trial-10	**Average**
アミノ酸	-	-	0.7750	0.9133	0.8700	**0.8535**	0.7461	0.8514	0.8266	0.6842	0.7647	0.8225	0.8107
糖鎖	-	-	0.7110	0.7131	0.8204	0.7296	0.6553	0.7069	0.5975	0.6760	0.6677	0.6904	0.6968
アンサンブルモデル	0.10	0.90	0.7110	0.9112	0.8700	0.8122	0.7523	0.8596	0.8369	0.6677	0.7709	0.8225	0.8014
	0.50	0.50	0.7110	**0.9298**	0.8741	0.7647	0.6821	0.8452	0.8225	0.6574	**0.8163**	0.8308	0.7934
	0.90	0.10	0.7110	0.8803	**0.8803**	0.7523	0.6677	0.8080	0.6182	0.6760	0.7812	**0.8617**	0.7637
結合モデル	-	-	**0.8369**	0.9009	0.8617	0.7792	**0.8349**	**0.8762**	**0.8906**	**0.7337**	0.7957	0.7523	**0.8262**

3.2. Human Protein Atlas

Human Protein Atlas(HPA)は，ヒトの様々な組織や癌細胞に存在するタンパク質の発現情報を集めたデータベースである．このうち，TCGA において遺伝子変異数を得た各糖タンパク質と対応する組織の発現量を抽出して，データセットを構築した．

3.3. GlyGen

GlyGen は，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である．GlyGen からは，遺伝子変異量が得られた糖タンパク質を対象として，糖鎖情報の抽出を行なった．

4. 実験

4.1. 実験設定

本実験では，遺伝子変異と関連のあるタンパク質をスクリーニングするユースケースを想定し，与えられたタンパク質のリストに対して，モデルが出力した遺伝子変異量に基づくタンパク質の順位予測を行うタスクを設定する．提案モデルの性能を評価するために，アミノ酸配列と糖鎖情報を組み合わせた(1)結合モデル，(2)アンサンブルモデルに加えて，(3)アミノ酸エンコーダモデル，(4)糖鎖エンコーダモデルからなる 2 つのベースラインモデルの計 4 つのモデルの性能の比較を行い，各入力情報の差異が性能に及ぼす影響を検証する．最終的に得られたデータセットである 89 サンプルの 70%を訓練用に，残りの 10%と 20%を開発用とテスト用にそれぞれ用いた．尚，トレーニングデータが少量であることを考慮し，モデルのトレーニングと評価においては，10-分割交差検定を行った．モデルのハイパーパラメータの調整には探索効率化ツールである Optuna[4]を使用した．

4.2. 評価指標

モデルの性能を評価するために，モデルが予測した遺伝子変異量に基づくタンパク質の順位と実際の順位との間でスピアマンの順位相関係数を計算する．また，性能比較においては，有意水準を 0.05 として統計的有意差検定を行う．

5. 結果と考察

5.1. 結果

表 1 に実験結果を示す．表 1 には，各エンコーダの出力に対する重み，各試行における順位相関係数及びその平均値を載せている．実験結果から，アミノ酸エンコーダモデルの方が，糖鎖エンコーダモデルよりも，順位相関係数の平均値が 10 ポイント以上高い結果となった．また，アミノ酸配列と糖鎖の組み合わせに関して，糖鎖とアミノ酸の結合モデルがすべてのモデルのうちで最も精度が高い結果となった．一方で，アンサンブルモデルによる精度は，アミノ酸エンコーダモデルと糖鎖エンコーダモデルの中間の結果となった．

5.2. 考察

実験結果は，糖タンパク質からの遺伝子変異量予測において，アミノ酸配列と糖鎖情報を組み合わせることが有効であることを示している．但し，独立した各エンコーダの予測に単純な重み付けを行うアンサンブルモデルの精度がアミノ酸エンコーダの精度よりも低かったことから，アミノ酸配列と糖鎖情報の組み合わせの方法が精度の向上に重要であるといえる．糖鎖エンコーダによる予測精度は最も低い結果となったものの，正の相関が得られたことから，糖鎖のみからの遺伝子変異予測の可能性も示唆された．

6. まとめ

本研究では，糖タンパク質から遺伝子変異量を予測するニューラルネットワークモデルの提案を行なった．実験結果から，タンパク質のアミノ酸配列に加えて，糖鎖情報が予測精度の向上に寄与することがわかった．今後の課題として，現在の小規模なデータセットよりも大規模なデータセットを構築して，モデルの性能を評価することが挙げられる．

謝辞

本研究は，AMED が助成する「糖鎖利用による革新的創薬技術開発事業」の支援を受けて行われたものです．また，本研究の遂行にあたり，ご協力いただいた創価大学木下研究室の塩田正明氏に感謝いたします．

参考文献

[1] Jin-xiao Liang, Yong Liang, Wei Gao. Clinicopathological and prognostic significance of sialyl Lewis X overexpresssion in patients with cancer: a meta-analysis. OncoTargets and Therapy, 9, 3113-3125, 2016.

[2] Masashi Tsubaki, Kentaro Tomii, Jun Sese. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. Bioinfomatics, 35, 309-318, 2019.

[3] Kenichi Tanaka, Kiyoko F Aoki-Kinoshita, Masaaki Kotera, Hiromichi Sawaki, Shinichiro Tsuchiya, Noriaki Fujita, Toshihide Shikanai, Masaki Kato, Shin Kawano, Issaku Yamada, Hisashi Narimatsu. WURCS: The Web3 Unique Representation of Carbohydrate Structures. Journal of Chemical Information and Modeling, 54(6), 1558-1566, 2014.

[4] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, Masanori Koyama. Optuna: A Next-generation Hyperparameter Optimization Framework. KDD, 2016.