一般口演 | 医療データ解析

一般口演11
機械学習

2021年11月20日(土) 09:10 〜 11:10 G会場 (2号館3階232+233)

[3-G-1-06] 化合物の類似性を用いた医薬品のグループ予測

*清水 充子[1]、竹村 匡正[1] （1. 兵庫県立大学大学院　応用情報科学研究科）
*Juko Shimizu[1], Tadamasa Takemura[1] （1. 兵庫県立大学大学院　応用情報科学研究科）

キーワード：KEGG Databese, Drug Group, Machine Learning

医薬品の開発は、化合物の性質を決定する化学構造を修飾することにより行われている。そこで作成された医薬品がどのような薬効や有害事象をもつのか予測することができれば、医薬品の開発プロセス全体が極めて効率的になることが期待できる。我々の研究チームでは、これらの考えから既存の医薬品の薬効および各医薬品の化学構造に類似した化合物のリストを用いることで、未知の医薬品（化合物）の薬効を類推できないか検討してきた。具体的には医薬品データベースを用いて市販されている全ての医薬品および薬効を抽出した上で、各医薬品の類似化合物データを用いて、教師あり機械学習を用いて予測を行った。結果は、同一薬効の医薬品が多い場合は十分な学習データが得られ、かつある程度予測可能な薬効があることが認められた。一方で、十分に学習できない医薬品しか存在しない薬効や、そもそもの薬効分類が薬の構造に依存するというよりは、リアルワールドでの用法（抗がん剤など）で体系化されている場合もあり、予測精度の低い薬効も多く見られた。 そこで今回は、化合物の構造情報と直接関係することが期待できる「機能的に類似した医薬品グループ」を対象として、各医薬品とその医薬品に類似する化合物のデータから、教師あり学習による医薬品グループの予測を試みた。データは、 KEGG MEDICUSにおける Drug（薬剤）、 Compound(化合物)、 Dgroup（医薬品グループ）のクラス（抗ウイルス薬や消化器系用薬など22クラス）を対象とし、類似する化合物の抽出には、 GenomeNetにおける SIMCOMPを利用した。教師あり機械学習にはサポートベクターマシンを利用した。結果として、十分に教師データが存在する医薬品グループについては、類似性のある化合物から、医薬品グループが予測できる可能性が示唆された。

©一般社団法人 日本医療情報学会

化合物の類似性を用いた医薬品グループ予測

清水 充子[*1*2]、岸本 和昌[*1*2]、中井 隆史[*2]、山下 晃平[*3]、佐藤 瑞月[*3]、梶田 朝佳[*3]、竹村 匡正[*2, *3]

*1 京都大学医学部附属病院、*2 兵庫県立大学大学院応用情報科学研究科、
*3 兵庫県立大学社会情報科学部

Prediction of Pharmaceuticals Groups using Compound Similality

Juko Shimizu[*1], Kazumasa Kishimoto[*2], Takashi Nakai[*3], Kohei Yamashita[*3], Mizuki Sato[*3], Asaka Kajita[*3], Tadamasa Takemura[*3]

*1 Kyoto University Hospital, *2 Graduate School of Applied Informatics University of Hyogo,
*3 Department of Social Informatics, University of Hyogo

Drug development is carried out by modifying the chemical structures that determine the properties of the compound. If it is possible to predict what kind of efficacy and adverse events the drug prepared there will have, it can be expected that the entire drug development process will be extremely efficient. Therefore, this time, we will target "functionally similar drug groups" that can be expected to be directly related to the structural information of compounds, and predict the drug group by supervised learning from the data of each drug and compounds similar to that drug. I tried. The data covers the Drug, Compound, and Classes of DGroup (22 classes such as antiviral drugs and gastrointestinal drugs) in KEGG MEDICUS is used to extract similar compounds. I used SIMCOMP in. A support vector machine was used for supervised machine learning. As a result, it was suggested that the drug group could be predicted from the similar compounds for the drug group for which there is sufficient teaching data.

Keywords: KEGG Database, Drug Group, Machine Learning

1. はじめに

医薬品の開発は、化合物の性質を決定する化学構造を修飾することにより行われている。そこで作成された医薬品がどのような薬効や有害事象をもつのか予測することができれば、医薬品の開発プロセス全体が極めて効率的になることが期待できる。我々の研究チームでは、これらの考えから既存の医薬品の薬効および各医薬品の化学構造に類似した化合物のリストを用いることで、未知の医薬品（化合物）の薬効を類推できないか検討してきた[1]。具体的には、京都大学科学研究所で整備されている KEGG（Kyoto Encyclopedia of Genes and Genomes）データベースにおける KEGG MEDICUS/KEGG Drug 医薬品データベースを用いて市販されている全ての医薬品および薬効を抽出した上で、各医薬品の類似化合物データを用いて、教師あり機械学習を用いて予測を行った。結果は、同一薬効の医薬品が多い場合は十分な学習データが得られ、かつある程度予測可能な薬効があることが認められた。一方で、十分に学習できない医薬品しか存在しない薬効や、そもそもの薬効分類が薬の構造に依存するというよりは、リアルワールドでの用法（抗がん剤など）で体系化されている場合もあり、予測精度の低い薬効も多く見られた。

一方、KEGG MEDICUS には各医薬品対しての様々な分類情報が収められており、添付文書の JAPIC コードや YJ コードなどの情報も含まれている。その中で、KEGG DGROUP は化学構造を含めた構造的・機能的に類似のグループがまとめられており、これらの情報を用いることで、化合物から医薬品のグループが予測可能かを検討した。

2. 方法

KEGG MEDICUS データベースより、医薬品グループデータ、医薬品データ、化合物データを取得して、医薬品グループの22クラスと医薬品を介した化合物データを用いて、深層学習（ニューラルネットワーク）を用いて予測することを試みた。

2.1 対象データ

今回対象とした KEGG MEDICUS は、日本および米国の医薬品添付文書を統合したリソースであり、ゲノムの情報と疾患との関連、医薬品の作用・副作用との関連、薬剤応答や薬剤耐性等のデータが集積されている。その中で、KEGG DRUG は、国内の全ての医療用医薬品および大衆薬、欧州の主要な医療用医薬品、米国の殆どの医療用医薬品 11714（2021 年 8 月 31 日現在）を一元的に集約したデータベースであり、「D 番号」を用いて、各薬剤の全てのデータが記載されている。その中には、治療標的（ターゲット）や薬物代謝など分子間相互作用ネットワーク等の情報も含まれている[2]。KEGG DGROUP は構造的・機能的に類似の医薬品グループとして、D 番号のグループとしてまとめられている。DGROUP は、「DG 番号」が振られているが、各 DG には、「クラス」と言われる 22 の医薬品グループが登録されているものがある。表1に DGROUP のクラスのエントリを示す。

表 1 KEGG DGROUP **のクラス**

抗寄生虫薬	貧血治療薬
糖尿病治療薬	トランスポーター阻害薬
精神神経系用薬	抗菌薬
トランスポーター基質薬	抗悪性腫瘍薬
抗アレルギー薬	鎮痛薬
トランスポーター誘導薬	代謝酵素基質薬
循環器系用薬	アデニル酸シクラーゼ活性化薬
その他	抗炎症薬
抗凝固薬	代謝酵素誘導薬
消化器系用薬	抗真菌薬
抗ウイルス薬	代謝酵素阻害薬

DGROUP は、2279 のエントリがあり、その中でひとつ以上のクラスを持つものは 1655 エントリであった。

KEGG COMPOUND は、代謝化合物その他生命現象に関与する生体内および生体外の低分子化合物のデータであり、「C 番号」で示されている。エントリー数は 17358 である。

2.2 SIMCOMP

KEGG では、SIMCOMP というシステムが実装されている[3]。SIMCOMP では、ある医薬品や化合物に対して、化合物としての骨格全体の類似度を計算する仕組みである。そのため、例えばある医薬品に対して、その医薬品の化学構造と類似した化合物の一覧を取得することができる。今回は SIMCOMP 上の類似性の評価尺度(0-1)において、0.6 以上の C を対象とした。

2.3 学習データの作成

これらの KEGG データベースより取得したデータによって、任意の医薬品(D)の医薬品グループ(DGROUP)を取得することが可能になり、DGROUP から 22 のクラスを取得することが可能になる。また、医薬品(D)より SIMCOMP を用いてその医薬品と類似する化合物(C)を取得することが可能になり、目的変数を 22 クラス、説明変数を化合物 C としてのデータを作成した。学習データの概念図を図1に示す。

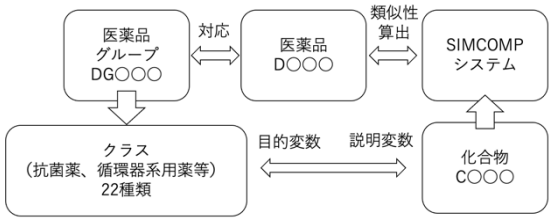

図 1　各データと学習データの関係

2.4 深層学習による分類

分類器は深層学習で構築した。環境構築としてプログラミング言語は Python 3.7、深層学習ライブラリは Keras 2.3.1(バックエンドは Tensorflow 2.0.0)を用いた。次にネットワーク設計は、中間を全結合 3 層で活性化関数 ReLU、出力層を 22 クラスで活性化関数 Sigmoid と損失関数 Binary crossentropy として 22 個の推定値がそれぞれ 0〜1 の値を出力するようにした。この推定値から閾値 0.5 を境にブーリアン型として各クラスを判定し、5 分割交差検証で評価した。

3. 結果

3.1 取得データ

22クラスを持つ医薬品(D)の数は、3791、出現した化合物(C)の数は 8623 であった。しかし、KEGG でエントリーされている 22 クラスのうち、トランスポーター誘導薬における医薬品(D)の数は4つ、貧血治療薬は6つ、アデニル酸シクラーゼ活性化薬は2つであり、これらのクラスについては学習対象から除外した。その他の各クラスのデータ数を表2に示す。

表 2　KEGG DGROUP のクラスごとの医薬品数

抗菌薬	493	トランスポーター阻害薬	62
抗真菌薬	55	精神神経系用薬	797
代謝酵素阻害薬	165	鎮痛薬	103
抗寄生虫薬	30	その他	718
代謝酵素基質薬	562	抗アレルギー薬	163
抗炎症薬	329	糖尿病治療薬	106
代謝酵素誘導薬	56	抗凝固薬	17
循環器系用薬	761	抗ウイルス薬	112
消化器系用薬	171	抗悪性腫瘍薬	312
トランスポーター基質薬	102		

3.2 予測実験結果

各クラスの予測結果(クラス名、精度、再現率、f1 値)を表3に示す。

表3　KEGG DGROUP の 19 クラスごとの予測結果

name	F1 score	Precision	Recall
抗菌薬	0.902386	0.969697	0.843813
抗真菌薬	0.9	1	0.818182
代謝酵素阻害薬	0.36214	0.564103	0.266667
抗寄生虫薬	0.545455	0.857143	0.4
代謝酵素基質薬	0.402612	0.518207	0.329181
抗炎症薬	0.854864	0.899329	0.81459
代謝酵素誘導薬	0.282051	0.5	0.196429
循環器系用薬	0.684251	0.835227	0.579501
消化器系用薬	0.5	0.709677	0.385965
トランスポーター基質薬	0.153846	0.6	0.088235
トランスポーター阻害薬	0.341463	0.7	0.225806
精神神経系用薬	0.726888	0.830645	0.646173
鎮痛薬	0.659218	0.776316	0.572816
その他	0.722222	0.809689	0.651811
抗アレルギー薬	0.526718	0.69697	0.423313
糖尿病治療薬	0.597403	0.958333	0.433962
抗凝固薬	0.714286	0.909091	0.588235
抗ウイルス薬	0.628571	0.873016	0.491071
抗悪性腫瘍薬	0.528634	0.84507	0.38461

4. 考察

今回、KEGG DGROUP におけるクラスに対して、各医薬品と構造的に類似の化合物を算出して、これらを説明変数として予測した結果、高い精度で予測できているクラスとそうでないクラスが見られた。その中でも、抗菌薬や抗真菌薬は高い精度で予測できており、化学構造と生体の作用点が明確であることが考えられる。

一方で、トランスポーター基質薬、トランスポーター阻害薬や、代謝酵素阻害薬、代謝酵素基質薬、代謝酵素誘導薬等の予測精度は低かった。これは、例えば、トランスポーター基質薬は降圧薬やスタチン系薬などがあり、作用点であるレセプター等の種類が多く、化学構造的に予測が難しいクラスと考えられるためと思われる。

参考文献

1) 清水充子他、医薬品情報データベースを利用した機械学習による医薬品の化学構造からの薬効推測に関する研究、第40回医療情報学連合大会論文集、2020

2) Minoru Kanehisa, Susumu Goto, Miho Furumichi, Mao Tanabe, Mika Hirakawa. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 2010 Jan;38

3) 田中伸也、五斗進.実践的 KEGG LIGLAND データベース活用法. 化学と生物. 44(5). p338-344. 2006.