3ZC-05

# 化合物の構造情報と非構造情報を用いたタンパク質ドッキング予測の為の機械学習手法の検討

伊東　忠佑†　　金盛　克俊‡　　大和田　勇人‡

東京理科大学大学院理工学研究科経営工学専攻†　　東京理科大学理工学部経営工学科‡

## 1. はじめに

現在，インシリコスクリーニングと呼ばれる創薬研究が盛んに行われている．中でも標的となるタンパク質と化合物との結合を予測することは創薬分野において重要である．

結合の予測には，ドッキングソフトが用いられるが計算時間が膨大にかかるため，近年，機械学習による手法が主流である．機械学習手法ではタンパク質のリガンド(結合化合物)とデコイ(非結合化合物)をトレーニングデータとして教師有り学習を行い，得られたモデルを使用してドッキング予測を行う．岡田らは化合物の物理的・化学的性質を商用ソフトウェアで計算し，得られた特徴を機械学習(SVM)のトレーニングデータとして用いている[1]．一方，Muggleton らは化合物の構造を用いた機械学習(ILP)が創薬に有効であることを示した[2]．これらは，化合物が持つ性質の 1 つのデータのみを利用している．

そこで，本論文ではこれらの学習法を併用し，非構造データ(物理的・化学的性質)と構造データ(化合物の構造)の 2 つのデータを学習したモデルの検討を行う．

## 2. 提案手法

本手法は次の 3 つの手順で構成される．

### 2.1 データ抽出

本手法では二つの機械学習を用い，それぞれ別々のデータセットを扱う．ILP では化合物の構造を表す構造データを学習し，SVM では化合物の物理化学的特性表す非構造データを学習する．

我々は，DUD-E から標的のたんぱく質に結合する化合物(リガンド)と結合しない化合物(デコイ)のデータ(mol2 ファイル)を取得する．

構造データは mol2 ファイルにテキスト処理を行い，ILP の入力形式に変換を行う．例えば，化合物の原子の構造として bond(C1,a1,a2,2)のような節がある．これは，C1(化合物)が a1(原子)と a2(原子)を持ち，それらが二重結合していることを表す．構造データはこのような論理式の集合で構成される．

非構造データは Discovery Studio を用いて 706 種の特徴を mol2 ファイルから計算する．しかし，すべての特徴を SVM で使用できるわけではない．例えば，SVM はある化合物の 1 つの特徴について 1 つの値のみ読むことができるが，1 つの特徴について複数の値を持つものが存在する．そのような特徴は利用することができない．加えて，すべての化合物で似たような値を持つ特徴は，機械学習に役立たない場合がある．このような理由から，学習に用いる特徴を選択する必要がある．

また，各特徴の範囲に大きな違いがあると，機械学習の性能は低下する．本手法では，各性質の範囲を正規分布ととらえ，範囲を標準正規分布にスケーリングする．

### 2.2 機械学習

データの抽出後，ILP と SVM を用いてそれぞれの結合予測モデルを作成する．

ILP は，帰納論理プログラミングとは，データの集合からデータを特徴づける論理的な規則を計算する枠組みである．ILP は一階述語論理を扱う事が出来る事から，属性値の集合では表現出来ない関係表現を学習する事が出来る．本手法では，以下の節を学習する．

bond(compound, atomid, atomid, bondtype)
atomobj(compound, atomid, atomtype, charge)
benzene(compound, benzeneid, atomid)

bond は原子同士の結合記述．atomobj は原子の種類の記述．benzene はベンゼンに含まれる原子の記述である．ILP はこれらの節を組み合わせ，リガンドのみに当てはまるルールを作り出す．例えば bond(A,B,C,2),atomobj(A,B,cl,minus_low),benzene(A,D,B)というルールがある．これは，化合物 A は二重結合で結ばれる原子 B と C を持ち，B は塩素で電荷は低くベンゼンに含まれるという意味である．ILP システムには溝口らが開発した GKS を用いる[3].

SVM はラベル(リガンド:1，デコイ:0)を目的変数として扱われ，特徴は説明変数として扱われる．これらの変数を利用して，SVM は二つのクラスを分ける超平面を求める．この超平面を用いることで，新しい化合物がリガンドに属するか，デコイに属するかを判別できる．SVM の判別性能はパラメータの cost と gamma に大きく依存する．そこでグリッドサーチを用いてパラメータを決定する．グリッドサーチは判別性能の指標に基づいて最良のパラメー

Investigation of a machine learning method for predicting protein docking using structured data and unstructured data of compounds
†Tadasuke Ito, Tokyo Univercity of Science

タを選択するものである．本実験のデータでは，デコイの数が多いためリガンドを検出しづらい．そのため，Recall と Precision の両方が高いときに，判別性能が高いといえる．本論文では Recall と Precision の調和平均をグリッドサーチでの評価指標とする．また，本研究では LIBSVM を使用する．

### 2.3 SVM と ILP の組み合わせ

構造を学習した ILP と非構造を学習した SVM の判別結果を組み合わせ，最終的な予測値を決定する．ある化合物に対して，予測値が一致した場合はその値を採択する．予測値が異なった場合，予測値の選択を行う必要がある．LIBSVM には予測値に対して，推定確率(確信の度合い)を算出する機能がある．そこでラベル 1 に対する推定確率によって，予測値を選択する．本手法では推定確率の閾値を 80%とする．推定確率が閾値を越える場合，SVM の結果を採択し，超えない場合は ILP の結果を採択する．

## 3. 評価実験

提案手法の有効性を示すため評価実験を行う．DUD-E にある 5 つたんぱく質を対象にした(表 1 ).データの量が少ないので，手法の評価には交差検証を用いる．

表 2 に実験結果を示す．ILP の分類精度は aofb を除いて高いものとなった．aofb ではリガンドをデコイと予測することが多くなり，recall が大きく下がった．これは学習をかける際，リガンドの数が少ないことが原因である．SVM の判別性能は全体的に高いものとなった．しかし，f1 を見ると aofb のみ 90%を下回った．ILP+SVM は aofb を除いて，他の手法より高い精度となり，f1 は 95%を超えるものとなった．このことから本手法を用いた高精度のスクリーニングが可能といえる．

ところで， ILP で構造を学習する利点として，分類ルールの視覚化が挙げられる．SVM では分類モデルがブラックボックスである．図 1 はこの実験で得られた thb の一部のリガンドが持つ共通ルールである． dock(A):-bond(A, B, C, 1), bond(A, D, E, 1), benzene(A, F, B), benzene(A, F, E), atomobj(A, C, cl, minus_low), atomobj(A, D, o, minus_low). 創薬研究者にとって，このように明確な判別モデルを得られることは重要である．

表 1. 実験データ

| Target Name | Description | Ligand | Decoy |
|---|---|---|---|
| aofb | Monoamine oxidase B | 122 | 366 |
| cah2 | Carbonic anhydrase II | 492 | 1476 |
| hs90a | Heat shock protein HSP 90-alpha | 88 | 264 |
| thb | Thyroid hormone receptor beta-1 | 103 | 309 |
| xiap | Inhibitor of apoptosis protein 3 | 100 | 300 |

表 2. 実験結果

ILP

| neme | accuracy | recall | precision | f1 |
|---|---|---|---|---|
| aofb | 0.824 | 0.369 | 0.833 | 0.511 |
| cah2 | 0.939 | 0.803 | 0.943 | 0.867 |
| hs90a | 0.949 | 0.898 | 0.898 | 0.898 |
| thb | 0.937 | 0.796 | 0.943 | 0.863 |
| xiap | 0.983 | 0.940 | 0.989 | 0.964 |

SVM

| neme | accuracy | recall | precision | f1 |
|---|---|---|---|---|
| aofb | 0.949 | 0.844 | 0.945 | 0.892 |
| cah2 | 0.981 | 0.959 | 0.965 | 0.962 |
| hs90a | 0.952 | 0.886 | 0.918 | 0.902 |
| thb | 0.951 | 0.874 | 0.928 | 0.900 |
| xiap | 0.983 | 0.970 | 0.960 | 0.965 |

ILP+SVM

| neme | accuracy | recall | precision | f1 |
|---|---|---|---|---|
| aofb | 0.924 | 0.746 | 0.938 | 0.831 |
| cah2 | 0.985 | 0.967 | 0.973 | 0.970 |
| hs90a | 0.974 | 0.943 | 0.954 | 0.949 |
| thb | 0.973 | 0.913 | 0.979 | 0.945 |
| xiap | 0.985 | 0.950 | 0.990 | 0.969 |

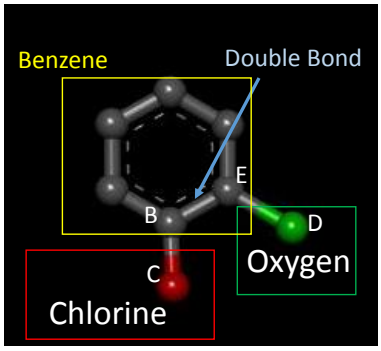


図 1. thb 共通ルール

## 4. 結論

本論文では構造データと非構造データを用いて機械学習を行い，化合物が標的タンパク質に結合するかどうかを分類する手法を提案した．5 つの化合物中 4 つで，本手法は単体での学習より高い精度となり，その全てで 97%を超え本手法の有効性を示すことができた．今後他のたんぱく質と化合物についても実験し検討を行っていく．

## 参考文献

[1] Okada M., Tsukamoto M., Ohwada H., Aoki S., Consensus Scoring to Improve the Predictive Power of in-silico Screening for Drug Design, Proc. of the 2nd International Conference on Engineering and Meta-Engineering, pp. 94-98, 2011.

[2] Muggleton S., Page D., Srinivasan A., An Initial Experiment into Stereochemistry-Based Drug Design Using Inductive Logic Programming., Lecture Notes in Computer Science Volume 1314, pp 23-40, 1997.

[3] Mizoguchi,F., Owada,H., Constrained Relative Least General Generalization for Inducing Constraint Logic Programs, New Generation Computing, Vol.13, pp.335-368, 1995.