4ZJ-02

タンパク質ーリガンド結合における ディープラーニングを用いた予測手法の開発

依田　洸[†]　　安尾　信明[†]　　関嶋　政和[†,‡]

東京工業大学情報理工学院情報工学系[†]
東京工業大学科学技術創成研究院スマート創薬研究ユニット[‡]

1. 序論

近年，医薬品開発におけるコストは年々増大しており，情報技術によるコストダウンの期待されておる．創薬の初期段階においては，多数の化合物が含まれるライブラリーから，計算機を用いて薬剤標的蛋白質に結合する化合物を発見するバーチャルスクリーニングのと呼ばれる手法として，タンパク質—リガンド間ドッキングが挙げられる．ドッキングにおける，様々なソフトが開発されているが，それらによる精度は未だ十分とは言えない[1]．

この問題を解決するため，ドッキングの結果に既知の化合物の評価結果を取り入れ，リランキングを行うことにより，精度向上を目指すSIFt[2]や Pharm-IF[3]などの試みが行われているが，タンパク質—リガンド間の相互作用における特徴量が必要である．

また，近年，機械学習のディープラーニング（深層学習）と呼ばれる分野の研究が盛んであり，画像認識においては，画像認識における精度を競う大会[4]を通じ，VGG や ResNet 等，様々なモデルが考案されるなど，著しい成果を上げている．深層学習の発展に伴い，Visual Inspection と呼ばれる人間の目によって判断することを深層学習を用いることにより，代替する試みが土木等の分野において研究が盛んに行われている．

バーチャルスクリーニングにおいて，ドッキング座標を用い，ディープラーニングを行う手法[5]などあるが，画像を用いた手法はない．

そこで，我々はドッキングの画像を用い，ディープラーニングをすることにより，既知の相互作用の特徴量を用いることなく，画像のみで活性の有無を判断するモデル，VisINet（VISual Inspection NETwork）を提案する．

また，ドッキングのベンチマークセットである DUD-E[6]を用いてバーチャルスクリーニングの精度評価を行った．

Protein-ligand binding prediction using deep learning
Hiroshi Yoda† Nobuaki Yasuo† Masakazu Sekijima†, ‡
†Department of Computer Science, Tokyo Institute of Technology
‡Advanced Computational Drug Discovery Unit, Tokyo Institute of Technology

2. 手法

2.1. 手法の概要

図の 1 に本研究におけるリランキング手法の概要を示す．本研究におけるリランキングでは，まず，a)標的タンパク質に対して活性が既知の化合物をドッキングした構造体情報を得る．次に b)ドッキングして得た構造体情報を用い，360°網羅的に画像化を行う．最後に c)ディープラーニングで画像を学習することにより，化合物の活性の有無を予測するモデルを作成する．活性が未知の化合物においては，同様に，a)ドッキングをし，b)画像化を行なったあと，d)生成されたモデルをベースとした VisINet により評価をし，e)リランキングを行う．最終的なスコアは，活性があるクラスに分類される確率として表される．

2.2. 機械学習

ドッキングをし，画像化された活性既知の化合物のラベルは，活性の有無を表す 2 値を用いた．ディープラーニングにおいては，50 層の ResNet である，ResNet-50 をベースとして改良を加えたものを用いた．学習においては，頑健性を持たせるため，入力画像として元の画像を 90° 180° 270° 360° ランダム回転したものを用いる．誤差関数としては，活性有無の逆比率を重みとした，重み付きクロスエントロピー誤差関数を用いた．勾配降下法として，Adam optimizer を用い，学習率はe^{-4}を用いた．

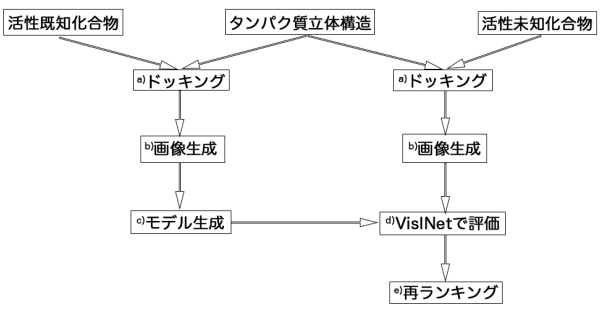

図 1:VisINet によるリランキングの概要

3. 実験

3.1. 実験条件

比較実験では，ドッキングに Glide version 65013，データセットに DUD-E を用いた．DUD-E には 102 タンパク質が登録されており，各タンパク質について共結晶構造，活性のある化合物，活性がない化合物が登録されている．本研究における実験では DUD-E のサブセットである Diverse subset 8 タンパクにおいて，共結晶構造のリガンドの位置に全化合物のドッキングをした．その後，得られたドッキング構造を，トレーニングセット，テストセットを 7:3 に分け，各々のデータセットにおいて，pymol を用い，各化合物 81 枚出力する 360° 網羅的画像生成を行なった．活性既知の化合物におけるドッキング画像が含まれるトレーニングセットにおいて，ResNet をベースとしたモデルを用い，複数ノード分散深層学習を行い，活性未知の化合物におけるドッキング画像が含まれるテストセットにおいて，VisINet により活性有無の評価した．評価値としては，上位 1%の化合物にどれだけ活性のある化合物が含まれるかを表す，EF1%，ROC 曲線下の面積である AUC を用いて Glide SP モードとの比較を行った．

3.2. 実験環境

実験は，東京工業大学のスーパーコンピュータ TSUBAME3.0 において行った．分散フレームワークは，Tensorflow，Horovod を使用した．4GPU を搭載した f ノードを複数ノード用い，実行した．詳しい実験環境について表 1 に示す．

表 1: 実験環境

CPU	Intel Xeon E5-2680 v4　2.4GHz × 2CPU
core/thread	14cores / 28threads x 2CPU
Memory	256GiB
OS	SUSE Linux Enterprise Server 12 SP2
GPU	NVIDIA TESLA P100 for NVlink-Optimized Servers x 4
SSD	2TB
Interconnect	Intel Omni-Path HFI 100Gbps x 4
Complier	gcc 4.8.5
MPI	openmpi/2.1.2
CUDA	CUDA 8.0.61
cuDNN	cuDNN 6.0

3.3. 結果

DUD-E Diverse Subset における 8 タンパク全てにおいて Glide SP モードの値を上回り，AUC，EF1%共に精度向上する結果となった．

4. 考察

VisINet を用いることにより，Glide SP モードに比べ，EF1%，AUC 共に精度向上が見られた．理由としては，主に二つ考えられる．

一つ目は，計算機性能の向上により，多量のデータを学習することが可能になったことである．本研究において，用いた画像データは，全体で 800 万枚を超え，膨大な量であった．以前では，計算機性能がボトルネックとなり，多量の画像を学習することが困難であったが，近年における計算機性能の向上に伴い，タンパク質-リガンドドッキングにおいても多量の画像を用い学習し評価することが可能となった．

二つ目は，評価関数の違いである．Glide SP モードにおいては，タンパク質-リガンド間における結合自由エネルギーを近似したものを評価関数として用いているが，タンパク質の自由度を無視し，剛体として扱っているため，結合自由エネルギーを正確に見積もることは困難であった．それに比べ，VisINet においては，画像を用いることにより，互いの原子の位置関係，結合ポケットの構造など様々な情報を学習し，評価することが可能となった．

本研究では，画像のみを用い，活性の有無が評価する手法である VisINet を提案した．VisINet は，Glide SP モードに対し，全てのタンパク質において，EF1%，AUC 共に精度が向上した．今後は，画像のどの部分が，活性評価に寄与するか解析を進めていきたい．

参考文献

[1] Lionta, E., Spyrou, G., Vassilatis, D. K. and Cournia, Z.: Structure-based virtual screening for drug discovery: Principles, applications and recent advances, *Current topics in medicinal chemistry*, **14**:16, 1923 (2014).

[2] Deng, Z., Chuaqui, C. and Singh, J.: Structural in- teraction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions, *Journal of medicinal chemistry*, **47**:2, 337–344 (2004).

[3] Sato, T., Honma, T. and Yokoyama, S.: Combining machine learning and pharmacophore-based interaction fingerprint for *in silico* screening, *Journal of chemical information and modeling*, **50**:1, 170–185 (2009).

[4] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton.:Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems.* (2012).

[5] Ragoza, Matthew, et al.:Protein‑Ligand scoring with Convolutional neural networks. *Journal of chemical information and modeling* **57**:4, 942-957 (2017).

[6] Mysinger, M. M., Carchia, M., Irwin, J. J. and Shoichet, B. K.: Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking, *Journal of medicinal chemistry*, **55**:14, 6582– 6594 (2012).