4B-07

深層学習に基づくタンパク質と化合物の相互作用予測

浜中 雅俊† 種石 慶‡ J. B. Brown† 奥野 恭史†
京都大学大学院医学研究科† 理化学研究所計算科学研究機構‡

1. はじめに

本稿では，医薬品となる化合物を発見するための第一段階のスクリーニングである，タンパク質と化合物の相互作用の予測について述べる．膨大な種類の化合物から医薬品になり得るリガンド化合物を見つけ出す工程は，開発にかかる時間とコストを押し上げる主要因となっている．

従来，タンパク質の立体構造と化合物との結合をドッキングシミュレーションで推定する研究が行われてきた[1]．しかし，多くの化合物では立体構造が未知であることや，予測的中率が低いという問題があった[2]．

我々はこれまで，相互作用が確認された 12.5 万件の結合データと，結合データに含まれない同数の組み合わせを非結合データとして用意し，それらをサポートベクターマシンで学習することで相互作用を予測する方法を提案してきたが，データが増えるにつれて学習時間が長大になることや，学習データが少数追加された場合でも再度学習をやりなおさなくてはならないなど，今後大規模な相互作用データを学習していく上で検討すべき課題があった[2]．

本稿では，相互作用予測に Deep Learning（深層学習）の一手法である，Deep Belief Networks (DBN) [3]を用いることを検討する．

2. タンパク質と化合物の相互作用予測

問題自体は単純な二値分類であるが，次のような特徴により新奇な化合物の予測は難しい．

ビッグスパースデータの取り扱い

化合物は 10^{60} 以上存在し，ヒトゲノムに２万種類以上のタンパク質がコーディングされているが実験により相互作用が既知なのはわずかである．また実験は，コストと時間の問題から特定の化合物群やタンパク質群においてなされる場合が多く，データが疎なエリアと密なエリアが存在する．統計的な手法を用いた予測では，密なエリアで予測精度が高く，疎なエリアで予測精度が下がる傾向にある．新奇な医薬品を開発するためにとって重要なのは，疎なエリアにある化合物であり，それらが医薬品となり得るかどうか高い精度で予測することは困難である．

Deep learning for estimating compound-proteins interactions
†Graduate School of Medicine, Kyoto University
‡Advanced Institute for Computational Science, RIKEN

人工的に生成した負例の影響

実験で相互作用が確認されたものは，後に論文や特許情報として公開されることで，データの収集が可能となるが，相互作用しないという結果を公開していることは稀で，収集できる負例の数は正例の数に比べて極めて少ない．そこで，人工的に負例を生成することで識別問題として扱えるようにしているが，負例の中には，誤ってラベル付けされている，すなわちもし仮に実験を行えば正例となるものが混ざっており，それが識別を難しくしている可能性がある．

本稿では前者の問題に対し，相互作用予測のための学習に DBN を用いることで，タンパク質および化合物から得られる特徴から重要な要素を残しながら低次元化していくことを目指す．重要な特徴が発見できれば，探索空間を大幅に縮小し，スパース性を解消できる可能性がある．

また後者に対し，DBN のプレトレーニングで，正例のみで学習を行うことを検討する．正例のみで同等の精度が実現できれば，学習の効率化が期待できる．また，誤ってラベル付けされた負例の影響を抑制し，識別率を向上させることも目指す．

2.1 相互作用データ

タンパク質のアミノ酸配列，化合物の化学構造の記述子をそれぞれ 1080 次元，990 次元で表現し，合計 2070 次元のベクトルとする．相互作用が確認されているタンパク質と化合物の組を，この 2070 次元のベクトルで表現したものを正例とする．正例に含まれないタンパク質と化合物の組を正例と同数用意してそれらを負例とする．

2.2 Deep Belief networks (DBN)

DBN は，プレトレーニングと呼ばれる教師なし学習を行うことで，これまで困難であった多層のニューラルネットワークの学習を可能としたものである．図 1 は我々が用いた DBN の構成を示したもので，入力はタンパク質および化合物の記述子から得た 2070 次元の特徴を平均 0 分散 1 で正規化した値，出力は相互作用がある場合に 1，相互作業がない場合に 0 を示す 1 次元である．

入力層と 1 段目の中間層および中間層とその次の中間層の間のネットワークは RBM (Restricted Boltzmann Machine)で構成し，教

師なし学習を行う．教師なし学習が終わった RBM の出力側の層と出力との間の教師つき学習は，ロジスティック回帰を使ったものや，サポートベクターマシンを用いたもの，ニューラルネットワークを用いたものなど複数の構成が考えられるが，本稿では，ニューラルネットワークで構成し，バックプロパゲーションを用いて最終段だけでなくネットワーク全体を学習させる．我々は，このように構築されたネットワークの重みを調べていくことで，今後，重要な特徴量を見つけることを目指している．

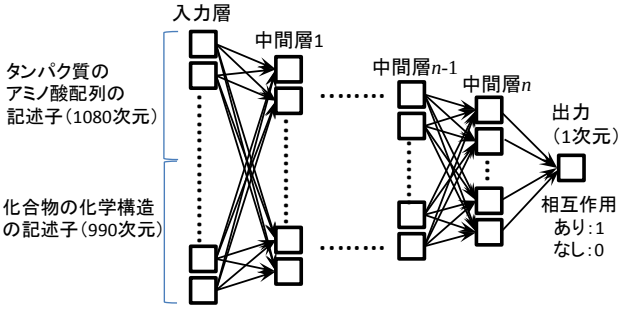

図 1: Deep Belief Networks

3. 実験結果

我々が整備しているデータベースから GPCR ファミリーの正例負例それぞれ 5000 件合計 1 万件をランダムで抽出し，5/6 を学習用データ 1/6 を評価用データとした [4]．DBN の構築には Accord.net ライブラリを使用し，ハイパーパラメータはライブラリの初期値を用いた．

3.1 教師なし学習で負例あり/なしの比較

教師なし学習で負例を用いる場合と用いない場合とを比較する．具体的には(a)正例 5000 件のみで教師なし学習を 1000 回行った場合と，(b) 正例負例あわせて 5000 件（2500 件ずつランダムで抽出）で教師なし学習を 1000 回行った場合で比較する．教師つき学習は，いずれも 1 万件で 1000 回行う．図 2 は，教師つき学習 50 回ごとにテストデータで評価した結果である．多くの学習回数で，正例のみのほうが高い性能であることが確認された．参考までに 1 万件で教師つき学習，教師なし学習を行った場合を図 2c に示す．精度は，正しく識別できたデータの数を評価データの総データ数で割ったものである．

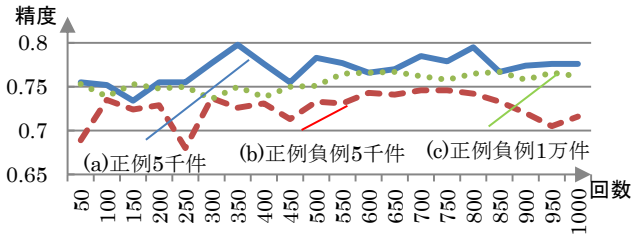

図 2: 負例あり/なしの性能比較

3.2 適切な中間層数・ユニット数の検討

中間層のユニット数が一定の割合で減少する構成に限定して適切な中間層数・ユニット数を検討する．たとえば中間層数 n が 2 でユニット数の倍率 j が 0.5 のネットワークでは，中間層 1 のユニット数は 1035（2070×0.5），中間層 2 のユニット数は，518（2070×0.5^2）となる．ユニット数の少数点以下は四捨五入とする．図 3 は，n を 1 から 7 まで，j を 0.3 から 1.0 まで 0.1 きざみで変化させたときの結果である．教師付き学習を 50 回ごとに評価したうちの最高値を示している．実験の結果，中間層数 6，倍率 0.6 の構成で精度が 0.805 となり最も高い性能であった．各層のユニット数はそれぞれ，1242，745，447，268，161，97 である．なお，教師なし学習は正例 5000 件，教師つき学習は 1 万件で行った．

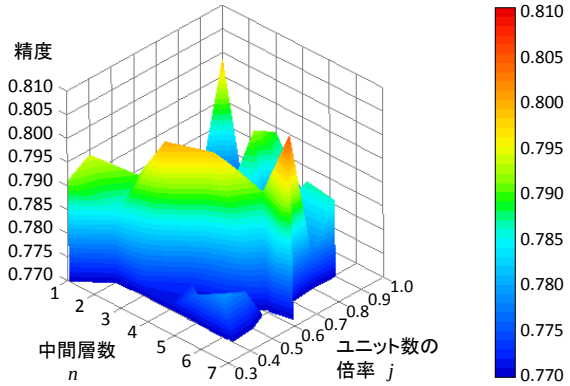

図 3：中間層数・ユニット数の検討

4. まとめ

本稿では，DBN を用いてタンパク質と化合物の相互作用予測を行い，負例なしのデータで効率的に教師なし学習ができること，中間層数が 6 でユニット数が層ごとに 0.6 倍になっていく構成の DBN で最も高い性能となることを確認した．今後，重要な特徴量の調査や，300 万件を超える大規模データでの評価を行っていく．

謝辞

本研究の一部は，科学技術振興機構(JST)の戦略的創造研究推進事業(CREST)の助成を受けた．

参考文献

[1] Fujitani,H. *et al.*: Massively parallel computation of absolute binding free energy with well-equilibrated states, Phys. Rev. E, 79, 021914, 2009.

[2] H. Yabuuchi *et al.*: Analysis of multiple compound-protein interactions reveals novel bioactive molecules, Mol. Syst. Biol., 7, p. 472, 2011.

[3] Hinton, G. E. *et al.*: A fast learning algorithm for deep belief nets, Neural computation, Vol. 18, No. 7, pp. 1527-1554, 2006.

[4] Okuno, Y. *et al.*: GLIDA: GPCR-Liand Database for Chemical Genomics Drug Discovery - Database and Tools Update, Nucleic Acids Research, 36, D907-12, 2008.