

# ディープラーニングを利用した タンパク質天然変性領域予測

相澤 洋輔<sup>1,2,a)</sup> 石田 貴士<sup>1,2,b)</sup> 秋山 泰<sup>1,2,c)</sup>

**概要:** 多くのタンパク質には特定の構造をとらない天然変性領域と呼ばれる区間が存在している。計算機を用いた機械学習により、タンパク質のアミノ酸配列をもとにした天然変性領域の様々な予測手法が提案されている。また、近年画像認識や音声認識で良い成果を上げたことで注目されている機械学習の手法として、ディープラーニングがある。ディープラーニングを利用したタンパク質の天然変性領域予測の手法としては、deep belief networks を利用した DNdisorder が提案されているが、その予測精度は最良の結果ではなかった。そこで本研究では、ディープラーニングにおけるもうひとつの手法である stacked denoising autoencoders を利用したタンパク質の天然変性領域予測を行った。提案手法の評価として、CASP9 データセットをテストデータとした予測性能評価を行い、ROC 曲線下面積で DNdisorder を超える性能を達成した。

**キーワード:** 天然変性領域, ディープラーニング, ニューラルネットワーク

## Protein Disorder Prediction Using Deep Networks

AIZAWA YOUSUKE<sup>1,2,a)</sup> ISHIDA TAKASHI<sup>1,2,b)</sup> AKIYAMA YUTAKA<sup>1,2,c)</sup>

**Abstract:** Many proteins include flexible regions generally called disordered region. Some methods for accurate prediction of disorder regions are proposed. We focus on deep learning because deep networks have already been successful predictors in the fields of image recognition and speech recognition. DNdisorder, which is the disorder predictor using deep learning with deep belief networks, was proposed, but its performance was not the best in the CASP9 evaluation. Thus, we tried to apply stacked denoising autoencoder, which is another method of deep learning, to the prediction of disordered protein regions. As a result, the proposed method is achieved performance of over DNdisorder in the CASP9 evaluation.

**Keywords:** disordered protein, deep learning, neural networks

### 1. はじめに

多くのタンパク質では特定の構造をとらない区間が存在し、その区間は天然変性領域 (disordered region) と呼ば

れている。天然変性領域は柔軟性を持ち、通常は水中を漂うひものように揺れ動いている。ただし、溶媒の変化や他のタンパク質との相互作用によって一時的に構造をとるものもある。天然変性領域はその性質上生物学において重要な意味を持ち、タンパク質の機能を同定するために重要な情報である [1]。さらに、近年ではこの柔軟性を利用してタンパク質間での相互作用を設計する disorder-based rational drug design (DBRDD) と呼ばれる手法も提案されている [2]。そのため、タンパク質の天然変性領域が今後の薬剤候補化合物探索の手法に大きく影響を与える可能性もある。

<sup>1</sup> 東京工業大学 大学院情報理工学研究科 計算工学専攻  
Graduated School of Information Science and Engineering,  
Tokyo Institute of Technology

<sup>2</sup> 東京工業大学 情報生命博士教育院  
Education Academy of Computational Life Sciences,  
Tokyo Institute of Technology

a) aizawa@bi.cs.titech.ac.jp

b) ishida@cs.titech.ac.jp

c) akiyama@cs.titech.ac.jp

タンパク質の天然変性領域を同定するための実験手法としては、X線結晶構造解析や核磁気共鳴分光法（NMR）などの手法が存在するが、生物学的な実験には時間と費用の面で大きなコストがかかるため、計算機による天然変性領域予測の手法が研究されている。一般的にタンパク質のアミノ酸配列は、その立体構造を決定する十分な情報を持っていることが知られており、タンパク質のアミノ酸配列から天然変性領域を予測する手法が提案されている [3]。しかし、依然としてその予測精度は改良の余地があり、さらなる向上が期待されている。

近年、画像認識 [4]、音声認識 [5] などの分野で高い性能を示したため、ディープラーニングが注目されるようになってきている。ディープラーニングは機械学習の手法のひとつであり、深い層構造のニューラルネットワークが利用される。局所解に陥る問題や学習時間の問題、計算量の問題により実際の応用はあまり進んでいなかったが、近年 Hinton らによる deep belief networks (DBN) [6] をはじめとする技術が提案され、局所解の問題を解決することで画像認識や音声認識の分野で良い成果を挙げ広く利用されるようになってきている。ディープラーニングの手法のひとつである DBN を利用したタンパク質の天然変性領域予測ソフトウェアとして、すでに DNdisorder [7] が開発されている。しかし、DNdisorder では DBN を弱学習器としたブースティングが行われており、CASP9 [8] データセットによる性能比較評価では、PrDOS2 [9] に劣る結果であった。その結果から、DBN が天然変性領域予測に適していない可能性や、チューニングが適切にされていなかった可能性が考えられた。

そこで本研究では、予測精度の向上のためディープラーニングを利用した天然変性領域予測を行った。学習にはディープラーニングの手法のひとつである stacked denoising autoencoders (SDA) [10] を利用し、ブースティングは行っていない。また、二次構造や溶媒接触性の情報を利用せずアミノ酸配列情報のみを入力特徴量とした。

## 2. 手法

本研究では、事前学習を利用したディープラーニングの手法である stacked denoising autoencoders (SDA) を用いてタンパク質の天然変性領域予測を行った。事前学習を利用する手法では、greedy layer-wise training とよばれる教師なし学習が行われる。1層目の事前学習が終わると結果はニューラルネットワークの初期値として決定され、その出力を2層目の入力として次の層の事前学習が行われる。それを繰り返すことにより、ネットワーク全体の初期値が1層ずつ順番に学習される。事前学習が全ての中間層で完了しニューラルネットワーク全体の初期値が決定すると、ラベル付き訓練データを利用して誤差逆伝播法により fine tuning とよばれる教師付き学習が行われる。

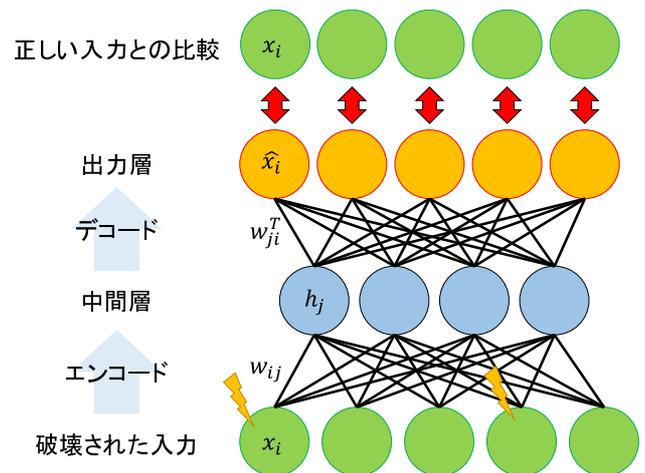


図 1 Denoising Autoencoder のネットワーク図

Fig. 1 Networks of Denoising Autoencoder

### 2.1 事前学習

SDA では、事前学習に denoising autoencoder (DA) [11] が利用される。DA では入力に対して確率的にノイズが加えられ入力の一部が破壊された後に、エンコード、デコードの処理が行われる。DA のモデルを図 1 に示す。稲妻マークがノイズによる確率的な入力の破壊を表している。

エンコードおよびデコードの活性の計算式は式 1 のように表される。

$$\begin{cases} h_j = \sigma \left( \sum_{i=1}^{N_x} C(x_i) w_{ij} + b_j \right) \\ \hat{x}_i = \sigma \left( \sum_{j=1}^{N_h} h_j w_{ji} + d_i \right) \end{cases} \quad (1)$$

ただし、 $x_i, h_j, \hat{x}_i$  はそれぞれ入力層、隠れ層、出力層のユニットの活性を、 $b_j, d_i$  はそれぞれ隠れ層、出力層のユニットのバイアスを、 $N_x, N_h$  は入力層と隠れ層のユニット数を、 $\sigma$  は活性化関数を、 $C$  は破壊関数を表す。本研究では、活性化関数にはエンコードに双曲線正接関数  $\sigma(x) = \tanh(x)$  を、デコードに線形関数  $\sigma(x) = x$  を利用した。また、破壊関数には確率的に入力が 0 に破壊されるごま塩ノイズを利用した。DA の学習において最小化する目的関数は  $x$  と  $\hat{x}$  の二乗誤差が用いられる。

### 2.2 Fine tuning

Fine tuning では、全体のニューラルネットワークにおいて誤差逆伝播法による教師付き学習を行う。本研究では汎化性能の向上のため、さらにミニバッチ方式の勾配降下法、慣性項、正則化項、学習率の減衰、早期停止を導入した。ミニバッチ方式勾配降下法は、標本をいくつか同じ個数ずつまとめたミニバッチを作成し、重みの更新の際にミニバッチ内で更新量を平均して学習を行う手法である。慣性項は、重みの更新の際に前回の更新量をある程度反映する手法である。正則化項には重みの L2 ノルムを利用した。

表 1 訓練用データセットの詳細  
Table 1 Detail of the training dataset

訓練用データセット	
タンパク質数	5,460
アミノ酸残基数	1,002,947
天然変性領域の残基数	67,285
天然変性領域の割合 (%)	6.71

学習率の減衰には、式 2 に示す焼きなまし減衰を用いた。

$$\eta = \eta_0 \frac{\tau}{\max(\tau, t)} \quad (2)$$

ただし、 $\eta$  は現在の学習率、 $\tau$  は減衰開始エポック、 $t$  は現在の訓練エポック数、 $\eta_0$  は学習率の初期値を表している。早期停止については、事前学習では 100 エポック学習を行い、fine tuning では 100 エポックの間確認用データセットの誤差が減少しなかった場合に終了するように設定した。

以上のような手法に伴うハイパーパラメータはディープラーニングの手法を導入する問題点として挙げられる。調整可能なパラメータは非常に多く、全探索を行うのは難しい。そこで、本研究では重要と考えられるパラメータのみ探索を行い、残りのパラメータは固定した。探索は学習率、慣性項、正則化項、学習率の減衰の順に行い、探索範囲はそれぞれ、 $10^{-1} \sim 10^{-5}$ 、 $0 \sim 0.9$ 、 $0.00005 \sim 0.0005$ 、 $0 \sim 100$  とした。訓練データセットにおける 3 分割交差検定の ROC 曲線下面積 (AUC) の値を基準に決定した。残りのパラメータについては Bengio による論文 [12] を参考に、ミニバッチサイズを 10、慣性項の係数は開始時に 0.06、250 エポック後に 0.2 に、正則化項の係数は 0.0003 に、学習率の初期値は事前学習で 0.001、fine tuning で 0.01 に、学習率の減衰開始エポックは 50 とした。

### 2.3 訓練用データセット

訓練用データセット (表 1) には、PrDOS2 の訓練で実際に使用されたものを利用した。これは、Protein Data Bank (PDB) に登録されているタンパク質からアミノ酸配列が完全に一致するタンパク質を除き、次の条件によって合致するタンパク質のアミノ酸配列から作成されたものである。

- X 線結晶構造において、解像度が  $2.0\text{\AA}$  以下である。
- X 線結晶構造における正確性を示す、R-factor が 0.25 以下である。
- アミノ酸配列の identity が 20% 以下である。
- アミノ酸配列の長さが 50 残基以上である。

また、タンパク質の PDB フォーマットのアミノ酸配列情報において、X 線結晶構造が missing residue と定義されたものが天然変性領域とされている。

タンパク質の天然変性領域はアミノ酸配列全体の中でも少ない割合で存在しているが、Hinton[13] によれば、ディープラーニングで学習を行う際には訓練データ内のクラスごとの割合、ミニバッチ内でのクラスごとの割合は等しいほ

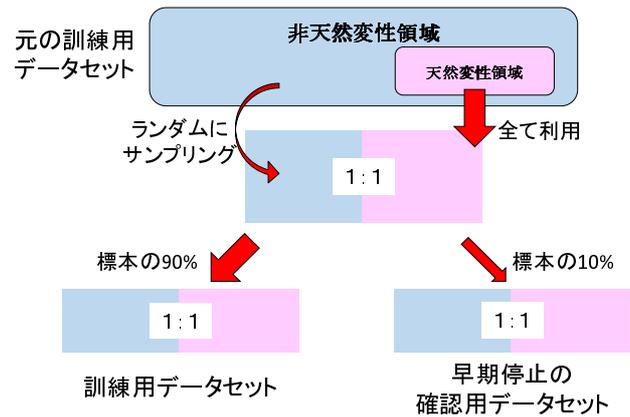


図 2 訓練用データセットの作成方法  
Fig. 2 Making of the training dataset

うが良いとされている。そこで本研究で用いる訓練データセットでは、訓練データ内の天然変性領域でないデータをランダムにサンプリングし、訓練データセットに元々含まれている天然変性領域のデータの数と合わせることで、それぞれのミニバッチ内に天然変性領域とそうでないアミノ酸残基が等しい割合で含まれるように調整を行った (図 2)。また、クラス均一化後の訓練用データセットから標本の 10% を早期停止のための確認用データセットとした。

### 2.4 入力特徴量

本研究では、PrDOS2[9] と同一の入力特徴量を利用した。これは、サイズ 27 残基のウィンドウ方式で特徴量を構成したもので、予測対象となる中心の 1 残基からアミノ酸鎖の両端に向かって 13 残基ずつのアミノ酸の情報が特徴量として含まれている。ウィンドウ内のアミノ酸残基の情報を表す特徴量には、タンパク質の進化情報を取り入れて予測するため位置特異的スコア行列 (position specific scoring matrix, PSSM) が利用されている。PSSM は NCBI の nr データベースに対する PSI-BLAST[14] の検索結果が用いられている。PSI-BLAST による類縁タンパク質の検索は、閾値となる E-Value スコアの値を 10、繰り返し回数を 2 回に設定して行われた。また、ウィンドウがタンパク質の N 末端または C 末端を超えてはみ出してしまった場合のために、1 ビットのスペーサー特徴量が導入されている。ウィンドウがアミノ酸配列に存在しない位置にあるアミノ酸残基情報を参照する際には、その位置の残基にあたる PSSM の特徴量が 0 となり、代わりにスペーサー特徴量が 1 となる。また、タンパク質の長さを表す 1 個の特徴量、中心の残基の種類を示す 20 個の特徴量が利用されている。以上の特徴量をまとめると、図 3 のようになる。入力特徴量の数は、ウィンドウ 1 つあたり 588 個である。

提案手法で利用する上で、PSSM の値は、 $[-10, 10]$  の範囲に切り捨て、 $[0, 1]$  の範囲に線形に正規化を行った。タンパク質のアミノ酸鎖の長さについては、500 残基以上の場

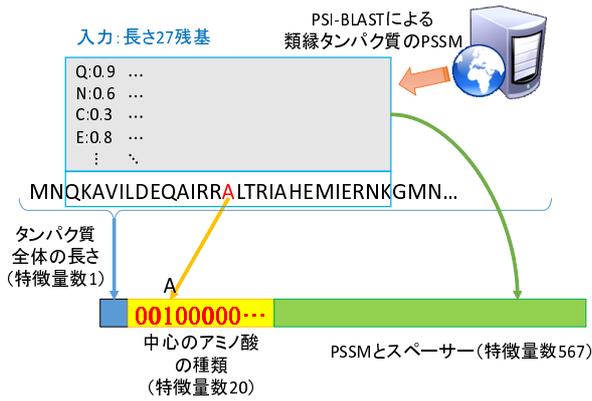


図 3 ウィンドウ方式での特徴量

Fig. 3 An example of generated features in the window

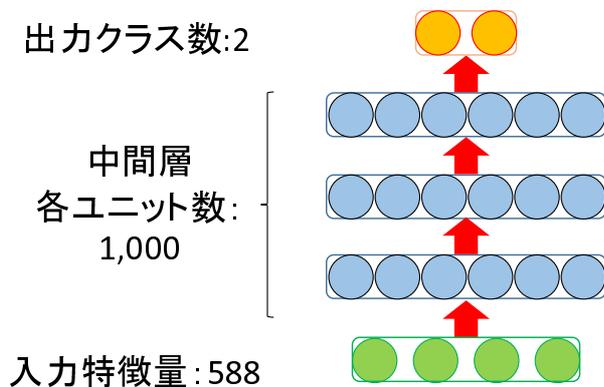


図 4 提案手法のネットワーク図

Fig. 4 Networks of the proposed method

合は値を 500 に切り捨て、 $[0,1]$  の範囲に線形に正規化を行った。

## 2.5 ネットワークの構造

SDA では、深い層構造のニューラルネットワークを利用する。本研究で利用したネットワークを図 4 に示す。中間層の数やユニット数はハイパーパラメータであり、対象となる問題に依存して最適な値は異なる。本研究ではハイパーパラメータ探索の簡略化のため、中間層は DNdisorder をはじめディープラーニングで一般的に用いられている 3 層に、ユニット数は入力特徴量に対して十分多いと考えられる 1,000 に設定した。

## 3. 評価実験

タンパク質の天然変性領域予測ソフトウェアとしては、多数の手法が提案されている。本研究では、その中でも特にディープラーニングを利用した手法である DNdisorder[7]、および CASP9 において最高の性能を示していた PrDOS2[9] との比較を行った。本研究で提案するネットワークと既存の 2 つの手法を比較するため、訓練用データセットおよび CASP9[8] データセットの 2 つに対して予測

表 2 CASP9 データセットの詳細

Table 2 Detail of CASP9 dataset

CASP9 データセット	
タンパク質数	117
アミノ酸残基数	26,083
天然変性領域の残基数	2,427
天然変性領域の割合 (%)	9.30

を行った。また、機械学習アルゴリズムとしての性能を評価するため、PrDOS2 の 1 段階目の予測で利用されているサポートベクターマシン (SVM) による予測結果も比較対象とした。

## 3.1 テストデータセット

### 3.1.1 訓練用データセット

CASP9 において最高の性能を示していた PrDOS2 は、その 1 段階目の予測手法に SVM を利用していた。それに対し、提案手法ではディープラーニングの手法のひとつである SDA を利用している。2 つの予測手法の違いによる性能の比較を行うため、訓練用データセットに対する 10 分割交差検定を行った。

訓練用データセットの詳細については、2.3 節で示したとおりである。SVM に関しては、石田ら [15] によりすでに 10 分割交差検定が行われている。本研究ではその論文に記述されている結果をもとに、ディープラーニングの手法の一つである SDA を利用した提案手法と、SVM を利用した手法の性能を比較した。

### 3.1.2 CASP9 データセット

2010 年に行われたタンパク質立体構造予測ベンチマークである CASP9 の天然変性領域予測で利用されたデータセット (表 2) に対して予測を行った。特徴量としてはタンパク質のアミノ酸配列のみが与えられ、ラベルとして天然変性領域に含まれるか否かの 2 値が紐付けられている。CASP9 には比較対象としている PrDOS2 が参加している。DNdisorder は参加していなかったものの、性能評価のため CASP9 データセットに対する予測を行っており、その論文中 [7] で結果を公開している。提案手法と DNdisorder、PrDOS2 および SVM の性能を比較評価するため、CASP9 をテストデータセットとする予測を行った。

## 3.2 評価方法

機械学習による予測の評価の手法としては Accuracy, F-measure など様々な評価指標が使われているが、天然変性領域の割合が少ない天然変性領域予測においては ROC 曲線下面積 (AUC) が一般的に用いられている [8]。本研究ではこの AUC を用いて予測手法を評価した。

表 3 訓練用データセットにおける 10 分割交差検定の結果

Table 3 Results of 10-fold cross validation in the training dataset

Predictor	AUC	
	Value	SE
Proposed	0.9075	0.0032
SVM[15]	0.881	n/a

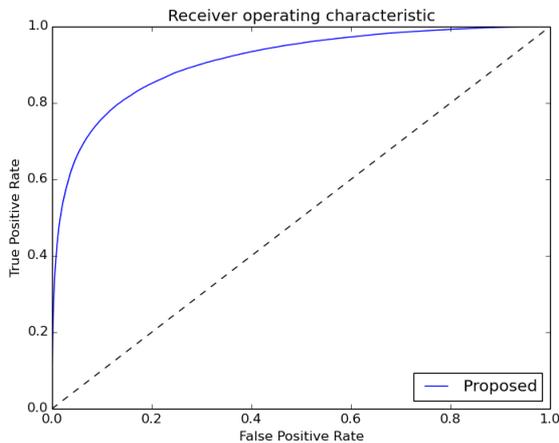


図 5 訓練用データセットにおける 10 分割交差検定の ROC 曲線  
 Fig. 5 ROC curve of 10-fold cross validation in the training dataset

### 3.3 実験結果

#### 3.3.1 訓練用データセットに対する 10 分割交差検定

訓練用データセットに対する 10 分割交差検定での提案手法の性能を表 3 に示す。SE は標準誤差を示している。また、提案手法の ROC 曲線を図 5 に示す。SVM による予測の AUC の値は、石田ら [15] により求められた値では 0.881 であった。それに対して提案手法の AUC は 0.91 程度となり、SVM に優る結果となった。

#### 3.3.2 CASP9 データセット

CASP9 データセットを予測対象とした場合の、提案手法と PrDOS2, DNdisorder および SVM の予測結果を表 4 に示す。ただし、PrDOS2[9] および DNdisorder[7] の性能は、論文に掲載されている値を引用している。95%信頼区間は、CASP9 における評価と同様に 1,000 回のギブスサンプリングにより算出している [8]。また、提案手法および SVM の ROC 曲線を図 6 に示す。AUC では PrDOS2 が 0.85 と最高値を示しているが、それに次いで提案手法と SVM は 0.84 程度のほぼ同等の性能となった。DNdisorder の AUC は 0.83 程度に留まり、提案手法が DNdisorder を超える性能を示した。また、ROC 曲線で比較すると、FP rate の低い部分では SVM が提案手法よりも若干良い性能を示した。

### 3.4 考察

提案手法は、CASP9 データセットに対する予測で DNdis-

表 4 CASP9 データセットに対する各予測器の予測結果

Table 4 Results of the prediction for CASP9

Predictor	AUC	
	Value	CI 95%
PrDOS2[9]	0.8544	±0.005
SVM	0.8409	±0.004
Proposed	0.8408	±0.004
DNdisorder[7]	0.8299	±0.005

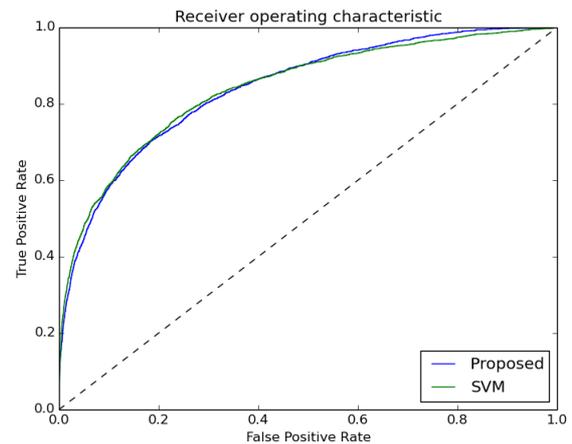


図 6 CASP9 データセット予測における提案手法の ROC 曲線  
 Fig. 6 ROC curve of the prediction for CASP9

order を超える性能を示した。DNdisorder の手法との違いとしては学習に DBN ではなく SDA を用いた点が挙げられる。DBN と SDA は問題に対して相性があり、天然変性領域予測に対しては SDA のほうが上手く学習が行えた可能性がある。しかし、入力とする特徴量や学習率の減衰などの細かい手法の違いもあるため、直接比較することは難しい。本研究で用いた学習率の減衰や訓練用データセットの均一化は一般的にディープラーニングに必要な手法であるとされており [12]、提案手法の性能の向上に寄与したと考えられる。

提案手法は、訓練データセットにおいて SVM を超える性能を示していたが、CASP9 データセットにおいてはほぼ同等の性能を示した。本研究では訓練用データセットにおける交差検定によりハイパーパラメータが設定されており、訓練データセットに対して過学習している可能性が考えられる。CASP9 データセット予測における ROC 曲線では、AUC はほぼ同等であったが FP rate の低い部分では SVMの方が若干良い性能を示していた。天然変性領域はタンパク質全体の中でも少ない割合で存在しているため、ROC 曲線においては原点付近での性能が重要と考えられる。その点では、提案手法が SVM に対してわずかに劣る結果となった。

CASP9 データセット予測においては、PrDOS2 が SVM を超えてトップの性能を示した。その理由としては、Pr-

DOS2 の予測手法では SVM による予測の後に相同性のあるタンパク質のアミノ酸配列による予測が行われる点が考えられる。提案手法は PrDOS2 に劣る結果となったが、SVM とほぼ同等の性能を示しており、提案手法においてもその後に PrDOS2 と同様の 2 段階目の予測を行うことでさらに性能が向上する可能性がある。

## 4. 結論

### 4.1 本研究のまとめ

本研究ではディープラーニングのひとつの手法である SDA を利用してタンパク質の天然変性領域予測を行った。また、高精度な予測を達成するために学習率の減衰、訓練用データセットのバランス化といった手法を取り入れ、ハイパーパラメータの調整を行った。本研究での提案手法の予測性能を評価するため、既存の予測器において最高性能を誇る PrDOS2、SDA と同じくディープラーニングの手法である DBN を予測に取り入れた DNdisorder の 2 つの手法と比較実験を行った。

結果として、国際的ベンチマークである CASP9 データセットに対する予測では AUC が DNdisorder を超える結果となった。本研究で行った学習率の減衰や訓練用データセットの均一化が精度の向上に寄与したと考えられる。

PrDOS2 に対しては、CASP9 データセットに対する予測結果において提案手法の性能が劣る結果となったが、PrDOS2 では SVM による予測だけでなく、その後に相同性のあるタンパク質のアミノ酸配列による予測を行っていた。SVM に対しては提案手法がほぼ同等の性能を示していたため、提案手法においても 2 段階の予測を行うことで精度が向上する可能性がある。

## 5. 今後の課題

本研究では、訓練用データセット内のクラス均一化のために非天然変性領域の標本の大部分を破棄していた。非天然変性領域の標本を全て利用しつつ、天然変性領域の標本をサンプリングにより水増しすることで性能が向上する可能性がある。

また、調整可能なハイパーパラメータの数はディープラーニングにおける大きな問題点とされており、効率的な手法は未だ開発されていない。本研究では、いくつかのハイパーパラメータを対象を絞ったパラメータ探索を行っており、より詳細なパラメータ探索を行うことで精度を改善できる可能性がある。

## 参考文献

[1] Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M., Obradovic, Z., Intrinsic disorder and protein function., *Biochemistry*, Vol. 41, pp. 6573–6582 (2002).  
[2] Cheng, Y., LeGall, T., Oldfield, C. J., Mueller, J. P., Van, Y. J., Romero, P., Cortese, M. S., Uversky, V. N.

and Dunker, A. K. Rational drug design via intrinsically disordered protein., *Trends in Biotechnology*, Vol. 24, No. 10, pp. 435–442 (2006).  
[3] Deng, X., Eickholt, J., Cheng, J., A comprehensive overview of computational protein disorder prediction methods., *Molecular BioSystems*, Vol. 8, pp. 114–121 (2012).  
[4] Le, Q. V., Ranzato, M. A., Devin, M., Corrado, G. S. and Ng, A. Y., Building High-level Features Using Large Scale Unsupervised Learning, *International Conference in Machine Learning* (2012).  
[5] Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. and Kingsbury, B., Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups., *IEEE Signal Processing Magazine*, Vol. 29, pp. 82–97 (2012).  
[6] Hinton, G., Osindero, S. and Teh, Y. W., A Fast Learning Algorithm for Deep Belief Nets., *Neural Computation*, Vol. 18, pp. 1527–1554 (2006).  
[7] Eickholt, J. and Cheng, J., DNdisorder: predicting protein disorder using boosting and deep networks., *BMC Bioinformatics*, Vol. 14, No. 88 (2013).  
[8] Monastyrskyy, B., Fidelis, K., Moulton, J., Tramontano, A. and Kryshtafovych, A., Evaluation of disorder predictions in CASP9., *Proteins: Structure, Function and Bioinformatics*, Vol. 79, pp. 107–118 (2011).  
[9] Ishida, T. and Kinoshita, K., PrDOS: Prediction of disordered protein regions from amino acid sequence., *Nucleic Acids Research*, Vol. 35(SUPPL.2), pp. W460–W464 (2007).  
[10] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y. and Manzagol, P., Stacked Denoising Autoencoders : Learning Useful Representations in a Deep Network with a Local Denoising Criterion., *Journal of Machine Learning Research*, Vol. 11, pp. 3371–3408 (2010).  
[11] Vincent, P., Larochelle, H., Bengio, Y. and Manzagol, P., Extracting and Composing Robust Features with Denoising Autoencoders., *Proceedings of the Twenty-fifth International Conference on Machine Learning*, pp. 1096–1103 (2008).  
[12] Bengio, Y., Practical recommendations for gradient-based training of deep architectures., *arXiv:1206.5533* (2012).  
[13] Hinton, G., A Practical Guide to Training Restricted Boltzmann Machines A Practical Guide to Training Restricted Boltzmann Machines. available from (<http://www.cs.toronto.edu/~hinton/absps/guideTR.pdf>) (2015.02.14).  
[14] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs., *Nucleic Acids Research*, Vol. 25, No. 17, pp. 338–402 (1997).  
[15] 石田貴士, 秋山泰, 単純ベイズ確率モデルの導入による天然変性蛋白質領域予測の改良, 情報処理学会研究報告バイオ情報学 (BIO), Vol. 21, No. 34, pp. 1–6 (2010).