2P1-15in1

表現学習と深層学習を用いたタンパク質の相同性検索と構造予測

Protein Homology Detection and Structure Prediction using Representation and Deep Learning

椿 真史 *1　新保 仁 *1　松本 裕治 *1
Masashi Tsubaki　Masashi Shimbo　Yuji Matsumoto

*1奈良先端科学技術大学院大学
Nara Institute of Science and Technology

Recent successful natural language processing (NLP) approaches are based on word representations and deep neural networks. On the other hand, in bioinformatics, 3D protein structure prediction from amino acid sequences is one of the important challenges. In this paper, inspired by recent machine learning techniques in NLP, we introduce a transition-based protein structure prediction using word representations and long short-term memory (LSTM), and evaluate the prediction using contact map. This transition-based approach, which uses polar coordinate transitions in 3D space of each amino acid over the protein sequence, allows us to naturally model contact map prediction compared to existing methods. In addition, we construct the model with LSTM which have the advantage of capturing long dependencies of amino acids in the protein sequence. We evaluated our method with the dataset of CASP11 which is a worldwide experiment for protein structure prediction.

1. 背景

タンパク質の相同性検索と 3 次元構造予測は，バイオインフォマティクスにおいて重要な問題である．相同性検索の目標は，与えられたタンパク質 (アミノ酸配列) を，データベースで決められた構造クラスに分類することである．また 3 次元構造予測の目標は，与えられたタンパク質中の全アミノ酸に対して，それらの 3 次元空間における位置を推定することである．バイオインフォマティクスではこれらの問題を，生物学の知識と機械学習の技術を用いて解くのが一般的である [Fariselli 99, Lena 12].

しかしながら，相同性検索のような大まかな構造クラス分類は高精度な一方で，3 次元構造予測の精度は未だ低いのが現状である [Lena 12]．また，予測に有効な配列や構造がデータベースに登録されていない場合，我々は与えられたアミノ酸配列のみからフルスクラッチでタンパク質の構造を予測しなければならない．これは *ab initio* モデリングと呼ばれ，物理化学や量子化学においても盛んに研究されている．

本論文で我々はまず，タンパク質を低次元ベクトルとして表現するために，近年の自然言語処理で成功した表現学習 [Pennington 14] を用いる．我々は，このベクトルがタンパク質の構造を適切に表現できていることを，簡単な構造クラス分類の実験で確認する．そして我々は，このベクトルを用いて，タンパク質の 3 次元構造予測を行う深層学習モデルを提案する．我々は提案法において，(1) タンパク質の 3 次元構造予測問題を，アミノ酸の系列データに対する 3 次元極座標の遷移予測問題として捉え直し (2) その際に重要な，配列中におけるアミノ酸間の長距離依存関係を，long short-term memory (LSTM) [Hochreiter 97] を用いて学習し (3) そして構造予測モデルと同時に，タンパク質ベクトル表現をさらに学習する．この一連の流れは，既存法のアプローチが持つ幾つかの問題を解決することができる．我々の知る限り提案法は，バイオインフォマティクスにおいて初めての，アミノ酸の極座標遷移に基づくタンパク質構造予測の深層学習モデルである．

連絡先: 椿真史，masashi-t@is.naist.jp

2. タンパク質データと既存研究の問題点

タンパク質の 3 次元構造情報は，PDB ファイルで提供されている *1．PDB ファイルには，タンパク質中の全アミノ酸の全原子について，それらの 3 次元空間における座標 (単位は $\AA$) が書き込まれている．我々は，この座標情報から原子間の距離などを計算することで，タンパク質の 3 次元構造を捉えることができる．

タンパク質の 3 次元構造予測には，様々な評価方法がある．特に本稿では，2 年に 1 度開かれるタンパク質構造予測の世界的なコンペティションである CASP*2 での評価方法の一つである，コンタクトマップを用いる．コンタクトマップとは，タンパク質中の全アミノ酸ペア *3 に対する 3 次元空間での距離情報を表現したマップ (正方行列) であり，特に配列中で 24 残基以上離れたアミノ酸間の距離が 8$\AA$ 以下の時，これをコンタクトと呼ぶ．コンタクトマップの評価には，全アミノ酸ペア (配列長が L の時，ペアの総数は $L(L-1)/2$) 中の上位 $L/5$ に対して，予測器がコンタクトと予測したものが実際にコンタクトである割合 (precision) を用い，これがコンタクト予測精度となる．

機械学習，特にニューラルネットワークを用いたコンタクト予測の研究は古くから存在し [Fariselli 99，Fariselli 01]，近年では深層学習を用いた手法も提案されている [Lena 12, Eickholt 12]．しかしながら，これらのアプローチの基本的な部分は以下の点で似通っており，またそれに伴い以下のような問題が生じる．

1. タンパク質中の個々のアミノ酸に対して，その生物学的な情報を特徴量として用いるが，これらは主にバイナリである．これにより，構造学習モデルの中で特徴量の再学習、つまり表現学習を行うことができない．

2. 既存法では主に，配列中の 2 つのアミノ酸に対する特徴ベクトルのペアを入力として，それがコンタクトか否か

∗1 http://www.wwpdb.org/
∗2 http://www.predictioncenter.org/
∗3 より正確には，各アミノ酸中の C_β 原子 (ただしグリシンには C_β 原子がないので C_α 原子) である．

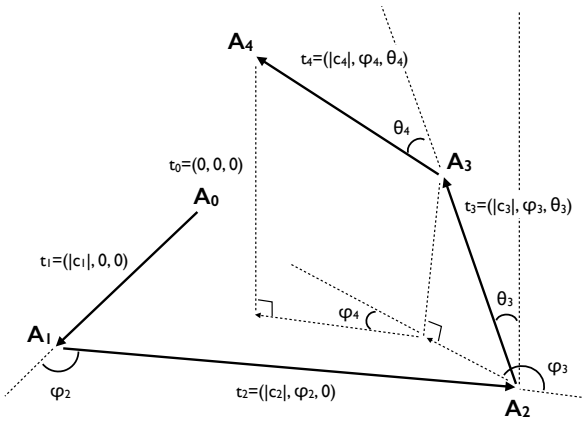

図 1: 極座標遷移の様子．A_i は i 番目のアミノ酸，$\mathbf{c}_i$ は PDB ファイルから計算される $i-1$ 番目から i 番目のアミノ酸への遷移ベクトル，ϕ_i と θ_i はそれぞれ $i-1$ 番目から i 番目への遷移ベクトルとのなす角，そして $\mathbf{t}_i$ はそれらの角を用いて表現した $i-1$ 番目から i 番目への極座標遷移である．

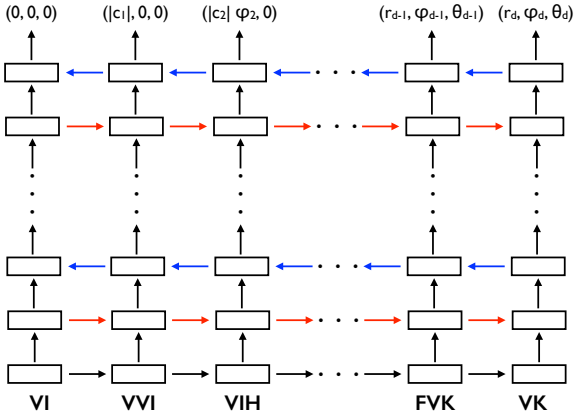

図 2: Stack bi-directional LSTM の概略図．前の層の出力を入力として双方向に LSTM を走らせた上で重ねていく．最終的な出力は各単語に対する極座標遷移であり，二乗誤差を最小化することで LSTM を学習する．また，開始と終了の 3-gram アミノ酸の先頭と末尾は空文字とする (“ VI” と “VK ”)．

を学習する．しかしこのアプローチは，配列の局所的な情報を個別に与えて学習しているに過ぎず，配列の大域的な情報を用いてタンパク質全体の構造を学習することはできない．

3. タンパク質中では一般に，2 つのアミノ酸がコンタクトである割合は非常に少ないため，既存法では学習の際に負例データのほとんどを捨ててしまう．

以上のような問題を我々は，次章で述べる提案法で解決する．

3. 提案法

3.1 タンパク質ベクトルを用いた構造クラス分類

我々はまず，約 90000 のタンパク質が登録されているデータベース CATH[*4] を用いて，3-gram アミノ酸を単語としたタンパク質コーパスを作成する．例えば，配列 “VVIHP” は，重複を許して “VVI VIH IHP” と単語分割される．次に，このコーパスに GloVe [Pennington 14] [*5] を適用し，各々の単語ベクトル表現を得る．そして，1 つのタンパク質ベクトルを，それに含まれる 3-gram アミノ酸ベクトルの総和として計算する．我々はこのタンパク質ベクトルを用いて，構造クラス分類の実験を行う．

3.2 アミノ酸の極座標遷移に基づくタンパク質の 3 次元構造予測

我々はまず，タンパク質を 3-gram アミノ酸を用いて分割して単語列を得る．次に，各々の単語に対して前述の単語ベクトルを割り当て，これを long short-term memory (LSTM) [Hochreiter 97] の入力とする．LSTM とは，系列データの長距離依存関係を適切に学習できる深層学習モデルの一つであり，自然言語処理において近年成功を収めている [Sutskever 14]．我々の提案法における LSTM の出力は，PDB ファイルから計算した各アミノ酸に対する極座標の遷移である．より具体的には，$i-1$ 番目のアミノ酸から i 番目のアミノ酸への極座標遷移 $\mathbf{t}_i = [r_i, \phi_i, \theta_i]$ であり，図 1 のように計算される 3 次元実数値ベクトルである．

提案法の LSTM の学習では，この 3 次元極座標遷移の系列を教師として，以下の二乗誤差

$$\mathcal{L}(\Theta) = \frac{1}{2}\sum_{d\in\mathcal{D}}\sum_{i=1}^{|d|}||\mathbf{t}_i - lstm(\mathbf{w}_i)||_2^2 + \frac{\lambda}{2}||\Theta||_2^2 \qquad (1)$$

を最小化する．ここで，$\mathcal{D}$ は学習するタンパク質のデータ集合，d はタンパク質，$|d|$ はそのアミノ酸配列の長さ，$\mathbf{t}_i \in \mathbb{R}^3$ は $i-1$ 番目の単語から i 番目の単語への極座標遷移，$\mathbf{w}_i \in \mathbb{R}^n$ は配列中の i 番目の n 次元単語ベクトル，$lstm(\mathbf{w})$ は単語ベクトルを入力として極座標遷移を出力する LSTM を表し，そして Θ はモデルの学習パラメータである．ここで Θ は，LSTM の重みパラメータと入力の単語ベクトル集合である．つまり我々は，§3.1 でコーパスから教師なしで学習した単語ベクトルを，教師ありの構造学習の中でさらに学習する．

このような極座標遷移を教師とした構造学習モデルは，§2 で述べた 3 つの問題を以下のように解決することができる．

1. 単語ベクトル表現は，構造学習の中でさらに学習することができ，3 次元構造予測により適した特徴量となる．

2. アミノ酸の特徴ベクトルペアを個別に与えて学習するのではなく，LSTM で配列中のアミノ酸間の長距離依存関係を考慮して，タンパク質全体の構造を学習できる．

3. タンパク質の構造をアミノ酸の系列データとして学習するため，大量の負例が存在するようなアンバランスデータとはならない．

また実験において我々は，LSTM の幾つかのバリエーションを用いる．特に、bi-directional LSTM [Graves 13] と stack bi-directional LSTM [Zhou 15] を用いる．提案法の bi-directional LSTM は，左から走らせた LSTM の出力を入力

*4 http://www.cathdb.info/
*5 http://nlp.stanford.edu/projects/glove/

手法	構造クラス分類精度 (%)
PSI-BLAST	47.5
GloVe	**63.2**

表 1: CATH の構造クラスである architecture の分類精度.

手法	コンタクト予測精度 (%)
LSTM	3.91
Bi-directional LSTM	5.08
3 layer stack bi-directional LSTM	**6.92**
6 layer stack bi-directional LSTM	4.87
[Eickholt 12]	**29.1**
[Lena 12]	**31.0**

表 2: CASP9 におけるコンタクト予測精度.

として，右からまた LSTM を走らせる．さらに，これを単純に積み重ねることで，stack bi-directional LSTM へと拡張する．Stack bi-directional LSTM の概略を図 2 に示す.

4. 実験

4.1 構造クラス分類

我々は，§3.1 で述べたタンパク質ベクトル (GloVe の次元は 100 とした) を用いて，CATH で定義されている構造クラスの分類を行う．特に本稿では，architecture という構造クラスの分類を行う．この構造クラスは，二次構造が折りたたまれた時にできる構造による分類であり，たとえ二次構造が異なっていても，折りたたまれた三次構造が似ているならば同じ architecture として分類される．データセットの作成には，CATH に登録されている約 90000 のタンパク質から，配列類似度の低いタンパク質 (BLAST の E-value が 10^{-6} 以上) をフィルタリングする．そしてその中で，データの多い順から 10 個の architecture を選ぶ．その結果，データセットのタンパク質は約 5000 となり，これを 4:1:5 に分割し，それぞれを訓練/開発/テストデータとする．実装には，scikit-learn[*6] のサポートベクターマシンを用い，カーネルはガウシアンカーネルとした．その他，ハイパーパラメータは開発データを用いてチューニングした.

表 1 に構造クラス分類の精度を示す．GloVe で得られたタンパク質ベクトルは，バイオインフォマティクスにおいてベースラインとして頻繁に用いられる PSI-BLAST よりも高い精度で，構造クラスを適切に分類できることが確認できる．これによりこのベクトルは，タンパク質の構造をある程度適切に表現できていると言える.

4.2 3 次元構造予測

我々は，訓練データに DNcon [Eickholt 12]，テストデータに CASP9 のデータセットを用いて，タンパク質の 3 次元構造予測モデルの学習とテストを行った．提案法の LSTM は Chainer[*7] で実装し，単語ベクトルは前述の 100 次元の GloVe，そして最適化には Adam [Kingma 14] を用いた.

表 2 にコンタクト予測の精度を示す．提案法の中では，3 層の stack bi-directional LSTM が最も良い精度であった．しかし，既存法の深層学習モデルの 2 つ [Eickholt 12, Lena 12] と比較すると，提案法の精度は著しく低い．既存法では，アミノ酸の生物学的な特徴量を，大規模なデータベースやバイオインフォマティクスのツール用いて作成しているため，深層学習モデルよりもむしろ，特徴量設計の時点で大きな差があると考えられる．提案法においても同様の特徴量を設計することで，性能を比較する必要がある.

*6 http://scikit-learn.org/stable/
*7 http://chainer.org/

5. 結論

本稿では，近年の自然言語処理において成功した表現学習と深層学習を用いて，アミノ酸の 3 次元極座標遷移に基づくタンパク質構造予測モデルを提案した．提案法の優位性を確認するには至らなかったものの，新たなアプローチの研究として進めていきたい.

参考文献

[Eickholt 12] Eickholt, J. and Cheng, J.: Predicting protein residue–residue contacts using deep networks and boosting, *Bioinformatics*, Vol. 28, No. 23, pp. 3066–3072 (2012)

[Fariselli 99] Fariselli, P. and Casadio, R.: A neural network based predictor of residue contacts in proteins, *Protein Engineering*, Vol. 12, No. 1, pp. 15–21 (1999)

[Fariselli 01] Fariselli, P., Olmea, O., Valencia, A., and Casadio, R.: Prediction of contact maps with neural networks and correlated mutations, *Protein engineering*, Vol. 14, No. 11, pp. 835–843 (2001)

[Graves 13] Graves, A., Jaitly, N., and Mohamed, A.-R.: Hybrid speech recognition with deep bidirectional LSTM, in *Automatic Speech Recognition and Understanding (ASRU)* (2013)

[Hochreiter 97] Hochreiter, S. and Schmidhuber, J.: Long short-term memory, *Neural computation*, Vol. 9, No. 8, pp. 1735–1780 (1997)

[Kingma 14] Kingma, D. and Ba, J.: Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014)

[Lena 12] Lena, P. D., Nagata, K., and Baldi, P. F.: Deep architectures for protein contact map prediction, *Bioinformatics*, Vol. 28, No. 19, pp. 2449–2457 (2012)

[Pennington 14] Pennington, J., Socher, R., and Manning, C.: Glove: Global Vectors for Word Representation, in *Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP)* (2014)

[Sutskever 14] Sutskever, I., Vinyals, O., and Le, Q. V.: Sequence to sequence learning with neural networks, in *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)* (2014)

[Zhou 15] Zhou, J. and Xu, W.: End-to-end learning of semantic role labeling using recurrent neural networks, in *Proceedings of the Conference on Association for Computational Linguistics (ACL)* (2015)