Vol.2019-MPS-124 No.4
2019/7/29

機械学習を用いたホモロジーモデリングのための配列アラインメント生成手法の高速化

成井政人[†1]　牧垣秀一朗[†1]　石田貴士[†1]

概要：ホモロジーモデリングは高精度な予測が可能なタンパク質立体構造予測手法であるが、その予測精度はテンプレート構造との配列アラインメントに大きく影響される。我々は教師あり機械学習の手法の一つである k 最近傍法をもちいることでこの配列アラインメントを改良する手法を提案し既存手法より質の高いアラインメントを生成することに成功した。しかし、その手法では典型的な長さのタンパク質 1 対の配列アラインメントに対して数万回の予測が必要となっており、その結果配列生成には数分から数十分の時間が必要となり、大規模な予測やデータベースへの検索などには不適となっていた。そこで本研究では配列アラインメントアルゴリズムと、予測アルゴリズムの改良を行い、配列アラインメント生成の高速化を試みた。その結果全体として約 21 倍の高速化を達成した。

キーワード：タンパク質構造予測、k 最近傍法、

Acceleration of machine learning-based sequence alignment generation for homology modeling

MASATO NARUI[†1]　SHUICHIRO MAKIGAKI[†1]
TAKASHI ISHIDA[†1]

Abstract: Homology modeling is a protein structure prediction method and practically useful because the accuracy of generated models is often high if we can find a good template structure. The method constructs a structure model based on a template structure and a sequence alignment between a query protein sequence and a protein sequence of the template protein. Thus, the quality of the sequence alignment is crucial for the prediction. Recently, a novel method to improve the sequence alignment using supervised machine learning technique was proposed. The method showed better accuracy compared with previous methods but required huge computation. Because almost millions of machine learning predictions are needed for a sequence alignment generation of a pair of protein sequences. Thus, in this study, we proposed a novel method to accelerate the sequence generation by optimizing the algorithm and parameters of machine learning predictions. As results, proposed method was approximately 21-times faster than the original one in exchange of trivial decrease in accuracy.

Keywords: Protein structure prediction, k-nearest neighbor method

1. はじめに

タンパク質は細胞内に存在する重要な生体高分子であり、様々な機能を有している。タンパク質はそのアミノ酸配列にしたがい、一定の立体構造を形成することが知られており、その機能は立体構造と深い関係性があることが知られている。そのため、タンパク質立体構造の決定は生物学的に重要な研究分野の一つとなっており、現在では X 線結晶構造解析や核磁気共鳴法などによって様々なタンパク質の立体構造が決定されている。しかし、実験的な手法による立体構造の決定には多くの時間とコストが必要となり、また結晶化の問題などで決定が難しいタンパク質が存在するため、古くは１９８０年代からアミノ酸配列情報を入力として計算機によりタンパク質立体構造を推定する手法が研究されてきた。

様々なタンパク質立体構造予測手法が提案されてきたが、現在のタンパク質立体構造予測手法は大きく２つの手法に分類される。１つは *de novo* 法と呼ばれる手法で主にタンパク質構造の自由エネルギーを最小化するように構造モデルを最適化することで予測を行う手法である。この手法はどんなタンパク質に対しても適用可能であるが、予測精度は現在でも不十分なものとなっている。もう一つの手法はホモロジーモデリングと総称される手法で、配列相同なタンパク質間ではしばしば構造が保存されていることを利用し、立体構造データベースに存在する構造既知のタンパク質を利用することで予測を行う手法である。ホモロジーモデリングにはテンプレート構造が見つからない場合は適用ができないという問題がある一方、良いテンプレートが見つかった際には非常に高い精度で立体構造が予測可能であることが知られている。しかし、構造ベース創薬などの応用を考えると、より高い精度の予測が望まれており、さらなる手法の改良が必要とされている。

ホモロジーモデリングの予測精度は予測対象となるタンパク質によく類似したタンパク質を探し出すテンプレート

†1 東京工業大学 情報理工学院
School of Computing, Tokyo Institute of Technology

構造探索に最も依存している。その一方、ホモロジーモデリングではテンプレート構造と同時に予測対象タンパク質とテンプレートタンパク質の配列アラインメントも予測の入力となっており、この配列アラインメントの質も最終的な予測モデルに大きな影響を与えることが知られている。そのため、多くの実用的な予測では予測精度の向上のために相同配列検索プログラムが出力する配列アラインメントをそのまま使用せずに、手動で修正することが行われてきた。

我々はこの問題を解決するために機械学習を利用した新たな手法を開発した[1]。図 1 に示すように、この手法は動的計画法によって配列アラインメントを計算する際に、その位置での一致度を、理想的なアラインメントである構造アラインメントを訓練セットとして機械学習を用いて構築された推定モデルを用いて推定することで、構造アラインメントに近い配列アラインメントを推定するものである。機械学習アルゴリズムとして k 最近傍法をもちいることで、この手法は TM-score[2]によるモデルの予測精度で平均 0.1 程度改善させることに成功している。しかし、この手法には一つの大きな欠点が存在している。この手法では、動的計画法のマトリックスサイズと同じ回数の k 最近傍法による予測が必要となっており、典型的な長さのタンパク質 1 対の配列アラインメントに対して数万回の予測が必要となる。その結果、提案時の実装では配列生成には数分から数十分の時間が必要となり、大規模な予測やデータベースへの検索などには不適となっている。

そこで本研究ではこのホモロジーモデリングのための配列アラインメント生成手法の高速化を試みた。まず、近似的な手法を用いることで k 最近傍法自体の高速化を試み、次に動的計画法による配列アラインの際に大きなスコアが得られそうなパス周辺以外では k 最近傍法による予測を行わないことで予測数の減少を図り、それによって高速化を行った。本研究での高速化では近似的な手法を導入したため、配列アラインメントが変わってしまう可能性があるが、評価実験では提案手法での配列アラインメントを用いて実際に構造の予測を行うことで、予測精度に与える影響を定量的に評価した。その結果、大きな予測精度の低下無しに配列アラインメントを含むタンパク質構造予測全体として 21 倍の高速化を達成した。

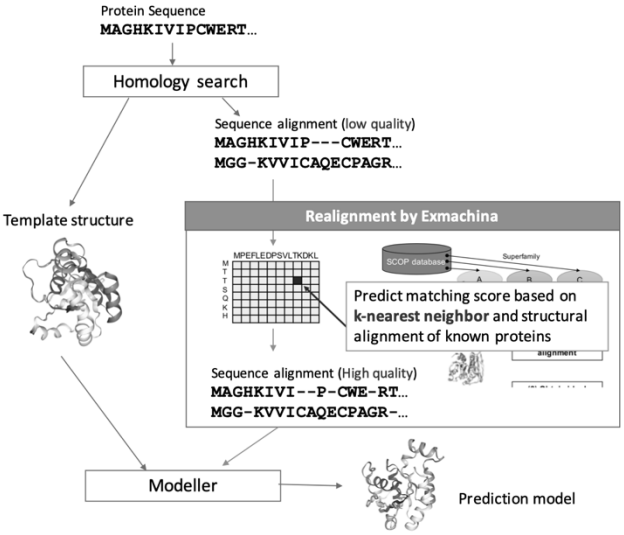

図 1　配列アラインメント生成の流れ
Figure 1　flow of sequence alignment generation.

2. 高速な k 最近傍アルゴリズムの検討

k 最近傍法の高速化には既存の高速化手法を複数検討し、立体構造予測手法の性質やデータセットに合ったものを適用することで実現した。既存手法では長さが M と N である 1 組のアミノ酸配列のアラインメントを生成するために MN 回の k 最近傍法による予測が必要である。このため、データベースに対し大量のクエリを計算する場合では、k 最近傍法の 予測回数は膨大なものになる。一方で、学習はデータベースに対して 1 度行えばよく、k 最近傍法の予測回数に対し 学習回数がかなり少なくなる。そのため、学習にかかる時 間は重視せず予測にかかる時間が短いアルゴリズムを特定した。また、検討の際にはホモロジーモデリングの精度を低下させない ために、近似近傍探索による精度低下が AUC 0.05 程度 に収まる範囲で行う。

2.1 kd-tree

今回のデータセットは 200 次元とそこまで高次元でないことや、木構造を用いるアルゴリズムで最も基本的なものであることを踏まえ、kd-tree[3]を検討した。

2.2 Randomized kd-tree

通常の kd-tree では分割する軸に分散が最大の軸を選ぶ ことが多い。しかし、多くのデータセットでは分散が最大 の軸と 2 番目以降の軸で分散に大きな差がないことが多く、最大の軸を選び続けることが最善の戦略とは限らない。この考えを元に、軸の選び方を変えた複数の木構造を用いるようにしたのが randomized kd-tree[4]である。

2.3 Hierarchical kMeans-tree

Muja らの研究 [5]より、近似による 精度低下が小さい範

囲では randomized kd-tree より hierarchical kMeans-tree が良い性能を示す可能性がある。 そのため hierarchical kMeans-tree[6]を実験対象に含めた。

2.4 実装

各アルゴリズムの実装は FLANN 1.9.1 を使用した。FLANN[15] は近似 k 最近傍法のライブラリである。k 最近傍法のライブラリは多数あるが、近似近傍探索が可能であることや、実装されているアルゴリズムの多彩さ、拡張しやすさから FLANN を採用した。

今回の実験では FLANN のソースコードから gcc コンパ イラでビルドしたものを使用した。ビルドにはライブラリの提供する cmake ファイルを用いた。コンパイルオプションも cmake ファイルに準じるが、C++コンパイラのオプションとして-std=c++11 を追加している。

2.5 ハイパーパラメータの決定とアルゴリズムの選択

FLANN の Python 実装を使用して scikit-learn 準拠の識別器モデルを作成した。この識別 器モデルを用いて scikit-learn の GridSearchCV にてハイ パーパラメータ探索を行った。また、構造予測を行い予測精度を推定するには多大な実行時間が必要となるため、k 最近傍法としての２状態予測の予測精度を精度として用いた。データセットはオリジナルの手法で使われているアミノ酸残基置換スコアを予測するデータセットを 1000 分の 1 にランダムサンプリングし たサブセットを用いる。

1000 分の 1 データセットに対して線形探索を行った結果は探索時間が 1,826 秒で AUC が 0.701 であった。この結果から、今回は AUC 0.05 の性能低下を許容し、AUC 0.695 を達成するまでにかかった探索時間が最も短いパラメータを最適なパラメータとした。

ハイパーパラメータの探索後、10 分の 1 にランダムサンプリングしたサブセットにより、最終的な予測精度と実行時間の評価を行った。結果を表 1 に示す。この実験では kd-tree に用いたものでは 1 日以内で実行が終了しなかったため、予測精度は不明となっている。最終的にあ Randomized kd-tree が最も高速で、予測精度の点でも良いバランスを示した。

表 1　アルゴリズム毎の実行時間と精度
Table 1　Prediction time and accuracy of k-nearest neighbor for larger test set

アルゴリズム	実行時間（秒）	精度（AUC）
Original (brute force)	47,226	0.721
Kd-tree	>86,240	N/A
Randomized kd-trees	998	0.716
hierarchical k-means tree	2,510	0.715

3. k 最近傍予測回数の削減

通常の配列アラインメントとは異なり、再アラインメントでは大まかなアラインメントのパスが推定可能である。そのため、その情報を利用することで動的計画法のマトリックスにおいて、対角線付近以外の箇所については最適パスが達しないと仮定することができる。そのため、パスの達しないと考えられる箇所についてはスコア計算を省くことで高速化を試みた。図２はこの高速化の概念図で、一般的なケースでは、k 最近傍法によって一致スコアを計算する範囲を４分の１まで縮小しても計算上問題ないことがわかる。

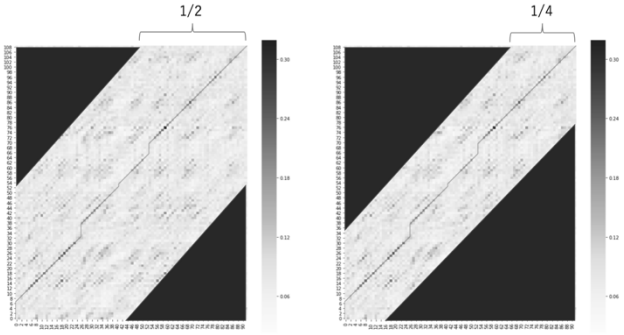

図 2　スコア計算範囲の削減
Figure 2　Reduction of matching score calculation.

3.1 アルゴリズム

動的計画法の各セルでの一致スコア$s_{x,y}$を下記の条件に該当する場合は、k 最近傍法により計算せずにに 0 で置き換える。

$$s_{x,y} = \begin{cases} 0 & \text{if } x > (1-r)x \text{ or } y > (1-r)y \\ s_{x,y}^{kNN} & \text{otherwise} \end{cases}$$

ここで、$s_{x,y}^{kNN}$ は k 最近傍法によって予測された一致スコアで、 r は省略する範囲を指定するパラメータである。小さいほど k 最近傍予測の回数が削減され、省略される範囲が大きくなる。図２に示されている通り、ほぼ最適パスが対角線に一致する場合は非常に小さい値を用いても結果は同じとなる。

4. 実験

表２は立体構造予測を行い、構造予測全体として Randomized kd-trees の導入による高速化と、予測削減による高速化がrを変化させたときにどう変化するかを示したものである。Randomized kd-trees の導入により実行時間は

６分の１程度まで高速化しているが、予測精度はほとんど低下していない。また、k 最近傍予測回数の削減による高速化ではrを 1/32 のように非常に小さなものとすると、大きな予測精度の低下が引き起こされている。特に 1/4 から 1/8 にかけて精度が大きく低下しており、現実的に利用可能な削減パラメータは 1/4 であると考えられる。このパラメータを用いた際の高速化は約 21 倍であるが、予測精度の低下は非常に小さなものであるため、有用であると考えられる

表 2　予測実行時間と立体構造予測精度

Table 2　Prediction time and accuracy of structure prediction

	kNN アルゴリズム	実行時間 (秒)	予測精度 (TM-score)
1	Original (brute force)	7,643	0.512
1	Randomized kd-trees	1,242	0.511
1/2	Randomized kd-trees	712	0.511
1/4	Randomized kd-trees	354	0.510
1/8	Randomized kd-trees	213	0.498
1/16	Randomized kd-trees	153	0.487
1/32	Randomized kd-trees	112	0.461

参考文献

[1] S. Makigaki and T. Ishida, “Sequence alignment using machine learning for accurate template-based protein structure prediction,” Bioinformatics, btz483, 2019.
[2] Y. Zhang and J. Skolnick, “Scoring function for automated assessment of protein structure template quality,” Proteins Struct. Funct. Genet., 2004.
[3] J. H. Freidman, J. L. Bentley, and R. A. Finkel, “An Algorithm for Finding Best Matches in Logarithmic Expected Time,” ACM Trans. Math. Softw., 2002
[4] R. H. Chanop Silpa-Anan, “Optimized KD-trees for image descriptor matching,” CVPR, 2008.
[5] M. Muja and D. Lowe, “Fast approximate nearest neighbors with autormatic alogrithm configuration,” in Proceedings of the Fourth International Conference on Computer Vision Theory and Applications, 2009.
[6] K. Mikolajczyk and J. Matas, “Improving descriptors for fast tree matching by optimal linear projection,” in Proceedings of the IEEE International Conference on Computer Vision, 2007.