Vol.2019-BIO-57 No.5
2019/3/8

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メント手法

牧垣 秀一朗[1,a)]　石田 貴士[1,b)]

概要：ホモロジーモデリングは相同性のある構造既知のタンパク質を鋳型に用いてアミノ酸配列からタンパク質の構造を予測する手法であり，鋳型となるタンパク質が発見できれば機能予測や薬剤開発でも使えるレベルの予測精度を達成できる手法である．しかし，構造未知のアミノ酸配列と既知の配列とのアライメントが最終的な予測結果の精度に大きく影響することが知られている．本研究では，アライメント生成時に用いるアミノ酸置換スコアを機械学習を用いた予測値に置き換えることで，より構造予測精度が向上するアライメントを生成する手法を示す．また，既存の相同性検索手法から得られるアライメントを用いた予測構造と比較して精度が向上することを確認した．

Sequence alignment method based on *k*-Nearest Neighbor for improving homology modeling

1. 序論

タンパク質は生物学や生化学，および製薬科学における重要な高分子であり，タンパク質の機能を明らかにするためには，タンパク質の構造と機能の関係を理解することが不可欠である．同様の機能を持つタンパク質はしばしば進化的に関連しており，これらのタンパク質はホモログと呼ばれている．タンパク質の機能を推定するために，これらの関連を明らかにし，その構造を研究することが分子生物学において重要である．タンパク質構造は X 線結晶学または核磁気共鳴などの実験的手段によって決定することができ，得られたタンパク質立体構造は，Protein Databank（PDB）に登録され，誰もが閲覧することができる．しかし，タンパク質構造を決定するための実験方法は改良されているにもかかわらず，アミノ酸配列を明らかにすることができる速度は，対応するタンパク質の構造を確かめる我々の能力を上回っている．したがって，タンパク質構造予測法，すなわち所与のアミノ酸配列の 3 次元構造モデルを生成するための計算技術の使用は，依然として不可欠である．

これまでタンパク質構造を予測するための様々な手法が提案されており，物理化学的シミュレーション（*de novo* とも呼ばれる）を用いた方法と，テンプレートベースまたはホモロジーモデリングと呼ばれる方法がある．後者は，テンプレートおよび予測対象アミノ酸配列とのアライメントに基づいて構造を予測する．鋳型となる構造は，相同性検索法で見いだされたホモログタンパク質の構造を用いる．予測対象タンパク質と構造的に類似した良いテンプレートと，予測対象配列とのアライメントを見つけることができれば，*de nove* 法と比べて予測モデルがはるかに正確であるため，現在，テンプレートベースのモデリング方法が最も実用的である．

古くは FASTA[2] および BLAST[3] による相同性検索研究が知られており，PSI-BLAST[4] および DELTA-BLAST のような，複数の配列アライメントに基づく配列プロファイル [5] を用いた手法が，高い精度で相同性を検出できている．また，配列プロファイルを用いた他の手法として，隠れマルコフモデルを用いた遠縁のホモログを検出する手法が存在し，HHpred[6] は構造予測ベンチマークで優れた性能を発揮している [7], [8].

昨今の相同性検索方法は遠隔のホモログを検出することができるようになってきているが，相同性検索プログラ

[1] 東京工業大学 情報理工学院
School of Computing, Tokyo Institute of Technology, Ookayama, Meguro-ku, Tokyo, 152–8550, Japan
[a)] makigaki@cb.cs.titech.ac.jp
[b)] ishida@c.titech.ac.jp

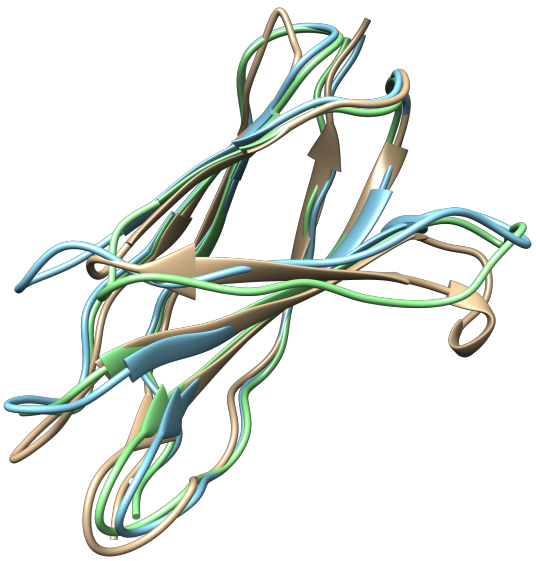

図 1　クエリ（黄）およびテンプレートタンパク質を，それぞれ 1QG3A および 1VA9A としたときの構造的差異．緑のモデルは構造アライメント（TM-align）から生成され，青のモデルは HHsearch の結果から生成されている．HHsearch と構造アライメントの TM-score は，それぞれ 0.801 と 0.881．（分子グラフィックスは UCSF Chimera[1] パッケージを使って生成した．）

Fig. 1　Model differences. Query (yellow) and template proteins are 1QG3A and 1VA9A, respectively. The green model is generated from structural alignment (TM-align), and the blue model is from HHsearch. TM-scores of HHsearch and structural alignment are 0.801 and 0.881 , respectively. (Molecular graphics was performed with the UCSF Chimera[1] package.)

ムによって生成された配列アライメントの質が悪いために，しばしば十分に正確な構造モデルを得ることができない [9], [10]．したがって，研究者は，より正確なモデルが必要な場合，品質を向上させるため，モデリングの前にアライメントを手動で編集する必要がある．多くの場合，相同性検索プログラムによって生成される配列アライメントは構造アライメントによって生成されるものとは異なり，特に遠隔ホモログについてはその違いが顕著である．構造アライメントにおいて予測対象タンパク質構造とテンプレートタンパク質構造との構造的差異は最小化されるため，構造アライメントによって生成された配列アライメントは，テンプレートベースのモデリングに理想的と言える（**図 1**）．このように，アライメントの品質はテンプレートベースのモデリングにとって非常に重要である．ホモロジーモデリングでは，テンプレートタンパク質なしで予測構造モデルを生成されることができないため，遠隔ホモログを検出することがより重要であると考えられてきた．しかし，より高精度のテンプレートベースのモデリングを達成するためには，配列アライメント生成の改善も重要である．最近では，機械学習法が相同性検出，フォールド認識，残基接触マップ予測，二面角予測，モデル品質評価，および二次構造予測の分野で力を発揮している（[11], [12], [13], [14], [15], [16]）．しかし，アライメント生成に関しては，それを分類問題または回帰問題として扱うことは困難であるため，この問題に関しては検討されていない．

本稿では，既知のホモログの構造アライメントを学習するような機械学習モデルに基づく，新しいペアワイズ配列アライメント生成法を提案する．機械学習を用いて配列アライメントを直接予測することは困難であるが，配列アライメント生成中の動的計画法実行時に，固定置換行列またはプロファイル比較を使用せずに，学習モデルから置換スコアを動的に予測する．この代替スコア予測プロセスにおいて機械学習が使用される．また，慎重に分割された訓練データセットとテストデータセットを使用して提案された方法を評価し，配列アライメント品質の尺度として，最先端の方法のものと予測構造モデルの精度を比較する．

2. 手法

一般的に，配列アライメント生成は相同性検出プロセスと統合されており，検出ツールはデータベースからの検索結果を用いて配列アライメントを出力する．本研究では，このアライメント生成にのみ焦点を当てているため，入力は予測対象アミノ酸配列（クエリ）および任意の相同性検索法によってテンプレートとして検出された別のアミノ酸配列であり，出力はホモロジーモデリングにより適したアライメントである．このプロセスはしばしば再アライメント（re-alignment）と呼ばれる．**図 2** はこのメソッドの概要を示している．古典的な動的計画法では，BLOSUM62 や PAM250 のような置換行列を用いて，残基ペア間の一致を評価する．アライメント精度を改善するために，FORTE[17] や FFAS[18] で使われているようなプロファイル比較法は，残基対の 2 つの位置特異的スコアマトリックス（PSSM）間の類似性を使用する．これらとは対照的に，本研究では教師あり機械学習に基づいて残基の一致を評価する．トレーニングデータセットのラベルとして，構造的に類似したタンパク質ペアのペアワイズ構造アライメントを使用して予測モデルを訓練する．従って，この方法は構造アライメントと類似した配列アライメントを出力すると期待される．提案手法は，入力としてクエリ配列とテンプレート配列を受け入れ，Smith-Waterman アルゴリズム [19] を使って配列アライメントを生成する．実際には，2 つの入力シーケンスは PSSM として表現され，固定サイズウィンドウ内の中心にある注目残基とその周囲の PSSM が使用される．最後に，配列アライメントとアライメントスコアを出力する．

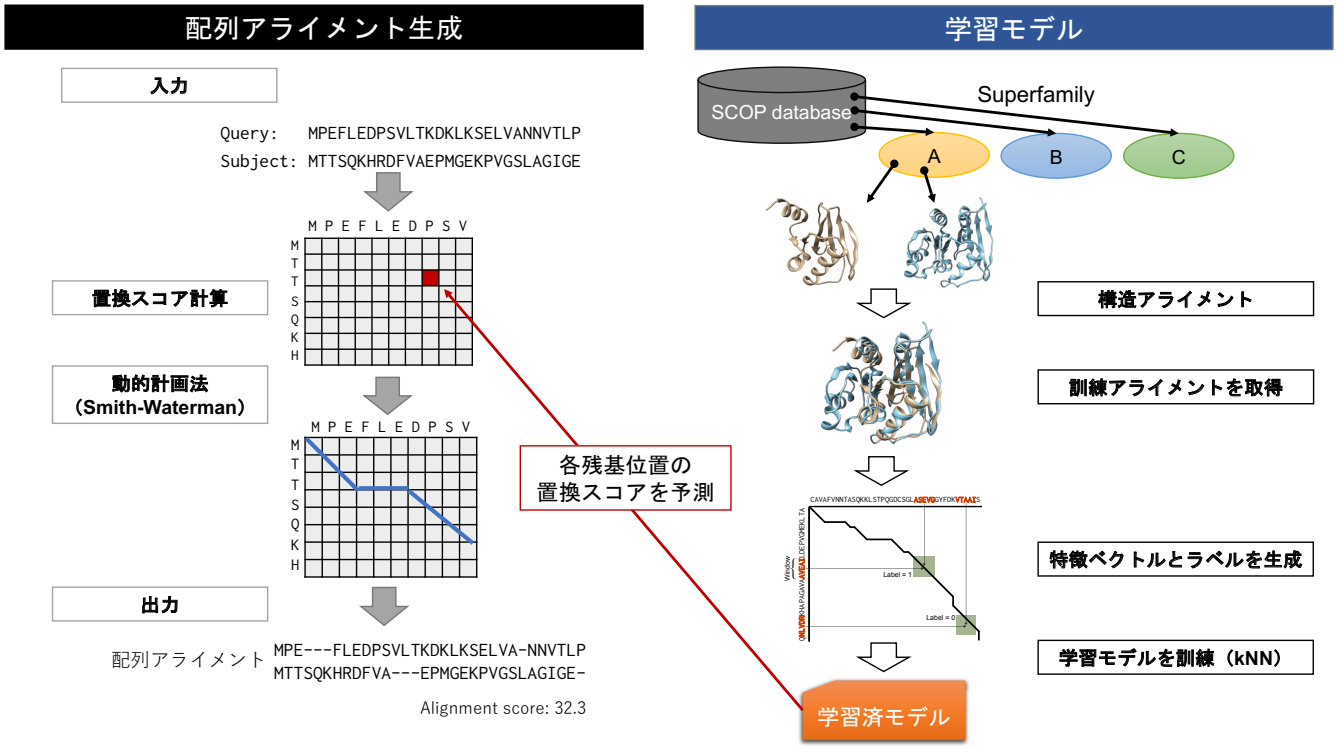

図 2　2 つの配列は Smith-Waterman アルゴリズムを使用してアライメントされ，このプロセスで使用される置換スコアは学習モデルによって直接与えられる．学習モデルは構造アライメントと同様のアライメントを出力するように訓練されている．

Fig. 2　Overview of the proposed method. Two sequences are aligned using the Smith–Waterman algorithm, and substitution scores used in the process are estimated by a prediction model. The prediction model is trained to output an alignment similar to the structural alignment.

2.1　データセット

我々の方法は既知の構造的に類似したタンパク質に関する情報を必要とし，本研究では SCOP（Structural Classification of Proteins）[20], [21] データベースを用いて構造アライメントを作成する．SCOP は，機能／構造分類に基づいて，タンパク質をクラス，フォールド，スーパーファミリー（SF），ファミリー，およびドメインに分類する．SCOP には冗長シーケンスが含まれており，過剰適合を避け，実行時間を短縮するために，代わりに SCOP40 を使用した．これは配列一致度が 40 % 未満のドメインのみを含むデータベースである．また，本研究では同じ SF にあるドメインを構造的に類似したタンパク質であると定義する．

正確な評価のために，テストデータセットを全データセットから分離する．本研究では，7 つの SCOP クラスからそれぞれ 2 つのドメインを選択し，様々なタンパク質構造タイプをカバーし，10 以上のドメインを含む SF からのみテストデータ用のドメインを選択した．小さい SF は無視し，残りのドメインを PDB の改訂日で並び替えた．最終的に**表 1** にリストされている 14 個のドメインがテストデータとして選択された．

トレーニングデータセットでは，TM-align[22] を使用して，同じ SF 内のすべてのドメインペアの構造アライメントを生成した．TM-align のスコア（TM-score[23]）が 0.5 未満であるドメインペアは構造類似度が低いペアとして扱い，それらは除外した [24]．また，SF にドメインが 1 つしかない場合もペアワイズアライメントが定義できいため無視する．この操作により 141 422 個のペアワイズ構造アライメントが生成された．この時に使った PSSM は，UniRef90 データベースに対し，PSI-BLAST を 3 回反復させることで生成した．トレーニングデータセットが合理的な計算時間内に処理するには大きすぎるようになったため，無作為選択によってその初期サイズの 1/10 に圧縮した．

2.2　特徴ベクトルとラベル

機械学習を使用してマッチングスコアを予測するには，残基ペアに関する情報を数値ベクトル表現でエンコードする必要がある．

(Q, T) をそれぞれクエリ配列とターゲット配列とし，Q_i を配列 Q の i 番目の残基とする．特徴ベクトル $\mathbf{V}_{x,y}$ は，クエリ配列とターゲット配列の特徴ベクトルの連結である．

$$\mathbf{V}_{x,y} = (\mathbf{P}_x^{query}, \mathbf{P}_y^{target}). \quad (1)$$

表 1　テストデータとして SCOP40 から 14 個のドメインを選択した．表示されているドメイン ID は SCOP の SID 番号．
Table 1　We selected 14 domains from SCOP40 as test data. Domain IDs shown are the SCOP sid numbers.

Class	Domains
a: All alpha proteins	d1wlqc_, d2axtu1
b: All beta proteins	d2zqna1, d1qg3a1
c: Alpha and beta proteins (a/b)	d1wzca1, d2dsta1
d: Alpha and beta proteins (a+b)	d1y5ha3, d2pzza1
e: Multidomain proteins	d1ni9a_, d3cw9a1
f: Membrane and cell surface proteins	d2axtd1, d2axto1
g: Small proteins	d2vy4a1, d3d9ta1

PSSM の形状は $20 \times N$ である．ここで 20 はアミノ酸種の数，N はタンパク質の長さである．$\mathbf{P}$ は PSSM 行の連結で，次のように定義される．

$$\mathbf{P}_i = (p_{i-\frac{w}{2}}, \ldots, p_i, \ldots, p_{i+\frac{w}{2}}), \tag{2}$$

ここで w はウィンドウの幅，p_i は PSSM の i 番目の行であり，i は以下によって制限されている．

$$x - \frac{w}{2} \leq i \leq x + \frac{w}{2}, \tag{3}$$
$$y - \frac{w}{2} \leq i \leq y + \frac{w}{2}. \tag{4}$$

$i \leq 0$ で定義される “パディング” 領域については，$|Q| > i$ と $|T| > i$，p_i を 0 に割り当てる．

ラベル $L_{x,y}$ を Q_x と T_y に割り当てて 0 または 1 とする．

$$L_{x,y} = \begin{cases} 1, & \text{if } Q_x \text{ matches } T_y \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

図 3 はこの特徴ベクトル生成の概要を示している．

2.3　アライメント生成

本研究では，ペアワイズ配列アライメントは Smith-Waterman アルゴリズム [25] を使用して計算されるが，このアルゴリズムには残基ペアの置換スコアが必要である．教師あり機械学習と上記で定義した特徴ベクトルを使用して，このスコアを予測する．具体的には，分類モデルの 1 つである k-最近傍（kNN）アルゴリズムを使用した．これは，特に大規模なトレーニングデータセットでは簡単で強力なためである [26]．kNN はバイナリラベル（0 または 1）を予測するが，バイナリラベルの予測はアライメント生成には粒度が粗すぎるため，kNN アルゴリズムの分類信頼スコアを Q_x と T_y の置換スコアとして使用する．

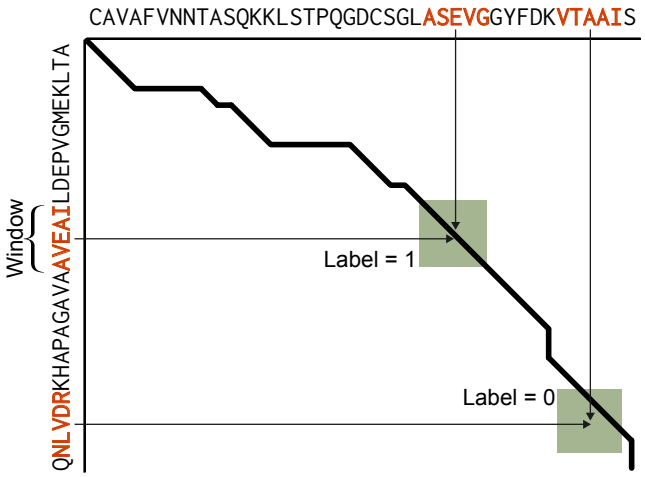

図 3　特徴ベクトル符号化方式の概要．X および Y 軸はアミノ酸配列を示す．黒の線は X 軸と Y 軸上の配列間の構造アライメントを示し，緑の長方形はウィンドウを示している．特徴ベクトルセットはこのウィンドウ内で計算される．特徴ベクトルはウィンドウ内配列の PSSM 列の連結である．注目残基がアライメントで一致する場合にラベルは 1 とし，それ以外の場合は 0 とする．
Fig. 3　Overview of a feature vector encoding scheme. The X and Y axes show an amino acid sequence. The green line shows the structural alignment between the sequences on the X and Y axes, with the green rectangle indicating the window. The feature vector set is calculated within this window. The feature vector is the concatenation of the PSSM columns of the window subsequence. If the current column is on the line, the label is 1; otherwise, it is 0.

2.4　パラメータ最適化

我々の方法にはいくつかのハイパーパラメータがあり，それらは 4 分割交差検証によって最適化した．交差検定については，210 のスーパーファミリーからテストデータ（> 10 のドメインを含む SF）と同じ基準で選択した．特徴ベクトルの符号化時のウィンドウ幅は 5（したがって特徴ベクトルの次元は 200），kNN の近傍の数は 1000，ギャップペナルティは gap-open に対して −0.1，gap-extend に対して −0.01 となった．提案手法の予測置換スコアは非常に小さい値となるため，これらのギャップペナルティは他の研究で使用されている一般的なギャップペナルティよりはるかに小さくなる．

3.　結果

我々の方法では 2 つの残基の一致スコアは kNN によって推定され，これらは 2 分類問題として扱うことができる．したがって，最初に ROC と AUC によるラベル予測プロセスの性能を評価し，その結果を**図 5** に示す．提案された方法は，d2axto1 を除いて，ラベルを正確に予測することができている．

次に，これらのアライメントから生成された予測タンパ

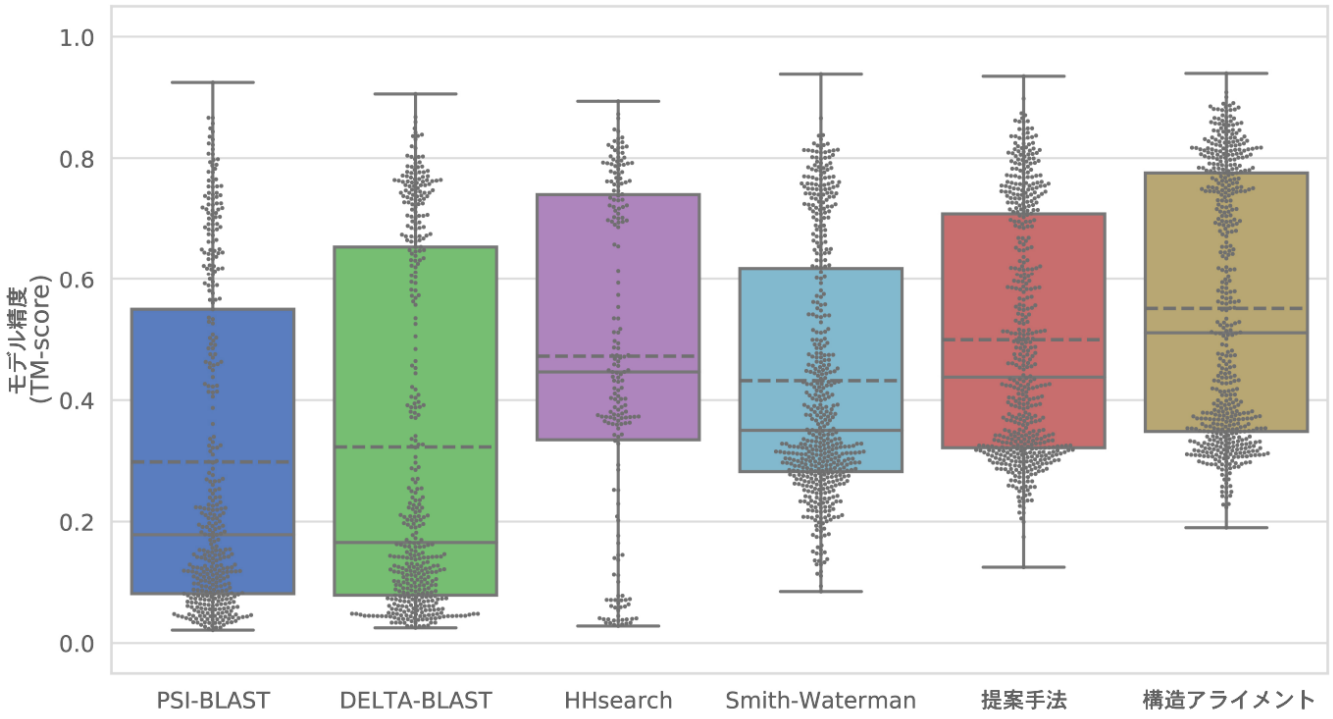

図 4 提案された方法および競合方法の配列非依存性 TM-score．実線は中央値を示し，点線は平均値を表す．点は各 TM-score におけるデータ密度を示す．

Fig. 4 Sequence-independent TM-score of the proposed and competitor methods. The solid line shows medians and the slashed line shows means. Dots indicate the data density at each TM-score.

ク質モデルの精度を比較して，生成された配列アライメントを評価した．これは，ラベル予測とモデルの正確性との間に強い相関関係がない可能性があるために必要であり，また，他の方法で直接比較することはできない．理論的には，構造アライメントに基づいて予測されたタンパク質モデルの精度は，単一テンプレートベースの予測において可能な最大の精度を示している．我々は，クエリタンパク質が属する SF に分類されている SCOP40 ドメインが，クエリと構造的に類似したタンパク質であると定義して，提案手法を適用した．モデルの正確さは，実験的に決定された構造と予測された構造との間の類似性を計算することによって評価することができるが，今回はその評価に TM-score[23] を使用した．これはモデルの正確さを 0.0 から 1.0 の範囲にスコアリングすることによってモデルの正確性を評価する．

まず，クエリタンパク質が属する SF 中の全てのドメインをテンプレートタンパク質として使用し，それらのペアワイズ構造アライメントを生成した．この構造アライメントの生成には TM-align[22] を使用した．次に，提案手法の精度を PSI-BLAST，DELTA-BLAST，HHsearch[27]，BLOSUM62 を用いた Smith-Waterman アルゴリズム，および構造アライメントの精度と比較した．配列プロファイルをクエリとして受け取る PSI-BLAST については，UniRef90 データベースで 3 回の PSI-BLAST 検索を実行してプロファイルを作成した．DELTA-BLAST は，検索の前に Conserved Domain Database[28] からプロファイルを検索するため，配列をクエリとして使用した．HHsearch では，クエリプロファイルを生成するために Uniclust20[29] を使用した．モデリングツールとして MODELLER[30] を使用した．

図 4 はタンパク質構造予測の精度を示している．予想されたように，構造アライメントは最も正確なモデル（平均で 0.551）を生成しており，提案された方法はほぼ同じ正確さ（0.499）を達成した．単純な Smith-Waterman アルゴリズムと HHsearch が 2 番目に精度がよく，それらの平均スコアはそれぞれ 0.432 と 0.472 だった．データ密度の結果から，提案手法を含むすべての方法で，結果は二峰性であった．すべての方法での最上位モデルは同様の精度を示したが，提案された方法を使用すると低い精度のモデル精度が向上した．

生成されたモデルの 1 つの例として**図 6** を挙げている．提案手法がモデルの精度，および実際のアライメントを改善できるていることを示している．アライメントの違いを示す**図 7** から，我々の方法は#46 と#55 の間の β シートターンを認識することができており，この部分がモデル精度の向上に貢献している．

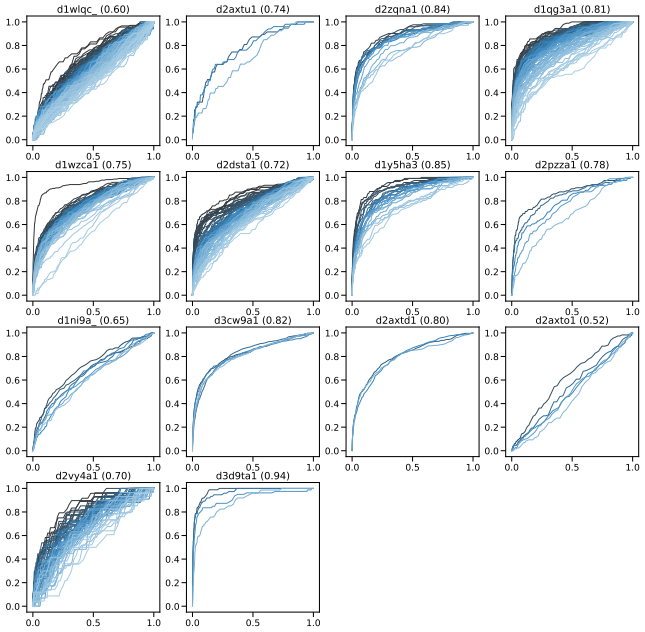

図 5 ラベル予測の ROC．タイトルは表 1 に示されている予測対象ドメイン名で，平均 AUC はその隣に示されている．

Fig. 5 ROC of label prediction. The title is target name shown in table 表 1, and average AUC is shown next to the name.

4. 考察

我々の方法は結果にアライメントスコアを含むため，そのアライメントスコアをソートすることによって相同性検出に使用することができる．この方法の相同性検出能力と，その検索結果の最上位アライメントを用いたモデルの精度を調べた．提案した方法の相同性検出性能を PSI/DELTA-BLAST および HHsearch の性能と比較した．合理的な計算時間のために，トレーニングデータセットは 1/10 ではなく 1/100 に圧縮したものを用いた．この評価では，クエリと同じスーパーファミリーのドメインを検出した場合を真，そして偽を異なる SF のドメインを検出した場合として定義した．結果を**表 2** に示した．ROC_n は n 番目までの偽陽性の結果のみを考慮し，AUROC_n は偽陽性の数とカットオフ n によって正規化された値である．PSI/DELTA-BLAST および HHsearch と比較して，提案した方法の検出感度は低かった．HHsearch の平均 AUROC_{50} は 0.690 だったが，提案された方法は最低のスコア 0.326 を示した．これは，提案手法が誤検出を多く示し，かつ長さによりスコアを正規化した E-value を本研究が採用していないためと考えられる．

このように，我々の方法を相同性検索に使用することはできないが，この検索結果の上位 10 件に対してテンプレートベースのモデリングを行い，3 次元モデルを作成した．モデルの精度は表 2 の 2 行目で示した．提案された方法は最高の平均 TM-score0.476 を達成した．また，DELTA-BLAST

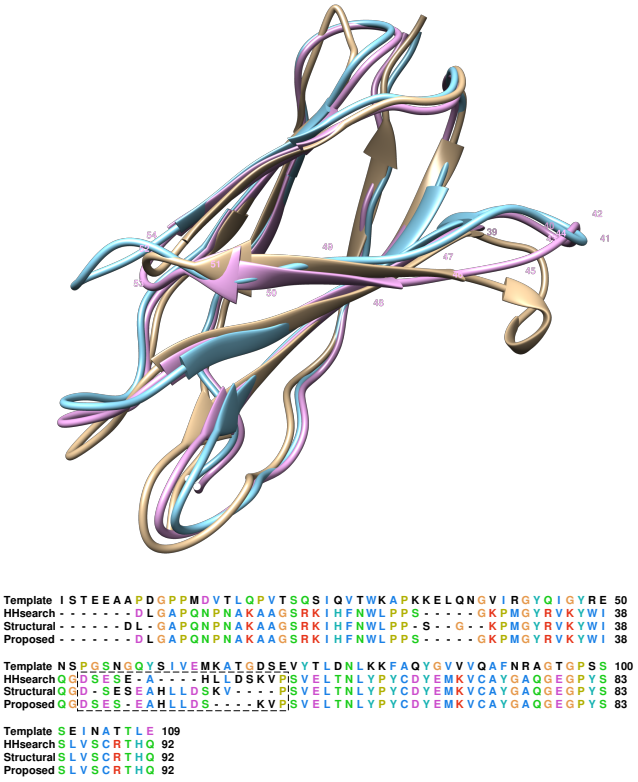

図 6 モデルの比較．黄のモデルは実験的に決定された構造を表している．赤いモデルは提案された方法によって生成され，青いモデルは HHsearch から生成されたモデルである．HHsearch と我々の方法の TM-score はそれぞれ 0.801 と 0.861 だった．

Fig. 6 Model comparison. The yellow model represents the native structure. The red model is generated the by proposed method, and the blue model is from HHsearch. The TM-scores of HHsearch and our method are 0.801 and 0.861 , respectively.

表 2 提案手法と競合手法の平均 AUROC_{50} とモデル精度（TM-score）

Table 2 Average AUROC_{50} and model accuracy (TM-score) of the proposed and competitor methods

	PSI-BLAST	DELTA-BLAST	HHsearch	Proposed
AUROC_{50}	0.435	0.491	0.690	0.326
TM-score	0.390	0.417	0.321	0.476

の結果は 2 番目に高く，0.417 だった．提案手法から生成されたモデルは他の手法から生成されたモデルよりも精度が高いことがわかった．これらの結果から，提案手法はテンプレート検出後のテンプレートベースモデリングのアライメント生成フェーズに有用であると考える．

5. 結論

本稿では機械学習を用いてタンパク質の構造を正確に予測する新しい配列アライメント生成法を提案した．固定されたアミノ酸置換行列の代わりに，提案手法は各残基対に

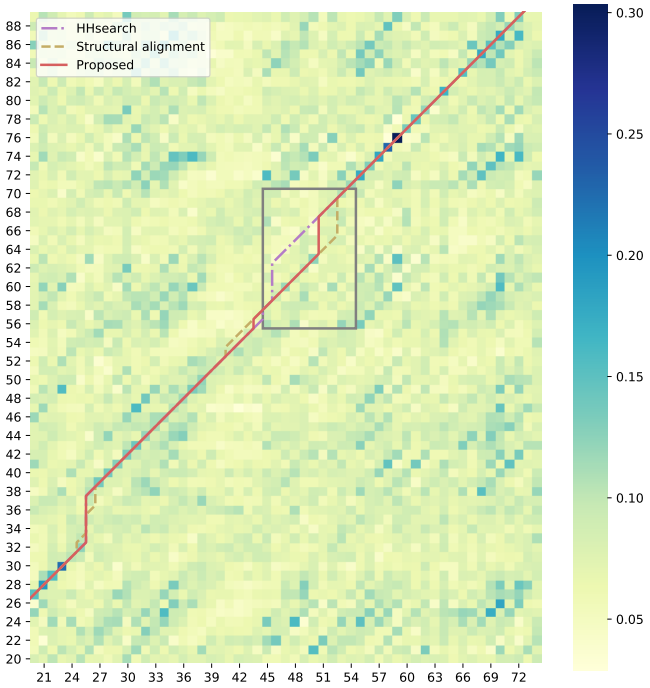

図 7 スコアヒートマップとアライメントパスの一部．X および Y 軸はクエリ（1QG3A）とテンプレート（1VA9A）の残基番号を示している．HHsearch（点線ダッシュ）は#46 と#55 の間で構造アライメント（ダッシュ）とは異なるアライメントを生成しているが，提案された方法（実線）は構造アライメントと同様のアライメントを生成している．

Fig. 7 Excerpt of score heatmap and alignment paths. X and Y axes shows query (1QG3A) and template (1VA9A) residue numbers, repeatedly. HHsearch (dotted dash) generated different alignment between #46 and #55 from structural alignment (dash) while proposed method (solid) could generate similar alignment to structural alignment.

おける置換スコアを動的に予測する．機械学習を適用するために，ペアワイズアライメントを潜在空間の数値ベクトルに変換する方法を開発した．これにより，教師あり機械学習アルゴリズムを使用することが可能になった．予測スコアはアライメントを生成するために直接使用され，その結果は次にテンプレートベースのモデリングのための入力として使用される．アライメント生成法をモデル精度により評価し，最先端の方法よりも優れていることを見出した．また，遠隔のホモログを検出する能力についても比較したが，AUROC$_{50}$ を比較すると，提案手法は他の方法よりも優れたパフォーマンスを示すことができなかった．しかし，提案手法は他の手法と比較して比較的正確な 3D モデルを生成することがわかった．

現在，提案手法は kNN とデータセットサイズのために長い実行時間が必要となっている．実行時間は予測対象タンパク質の数とその大きさに依存するため，本研究では使用するトレーニングデータの量を減らすことになった．この点において，我々の方法は厳密解を必要としないため，近似スキームを含む，より高速な kNN アルゴリズムを採用することは自然な拡張である．また，提案された特徴ベクトル設計は二次元として扱うこともできるため，将来的には，畳み込みニューラルネットワークなどの高性能モデルの使用も検討する．

参考文献

[1] Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. and Ferrin, T. E.: UCSF Chimera—A Visualization System for Exploratory Research and Analysis, *J Comput Chem*, (online), DOI: 10.1002/jcc.20084 (2004).

[2] Pearson, W. R. and Lipman, D. J.: Improved tools for biological sequence comparison., *Proceedings of the National Academy of Sciences*, (online), DOI: 10.1073/pnas.85.8.2444 (1988).

[3] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J.: Basic local alignment search tool., *Journal of molecular biology*, Vol. 215, No. 3, pp. 403–10 (online), DOI: 10.1016/S0022-2836(05)80360-2 (1990).

[4] Altschul, S. F., Madden, T. L., Schäffer, 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. 3389–3402 (online), DOI: 10.1093/nar/25.17.3389 (1997).

[5] Boratyn, G. M., Schäffer, A. A., Agarwala, R., Altschul, S. F., Lipman, D. J. and Madden, T. L.: Domain enhanced lookup time accelerated BLAST, *Biology Direct*, Vol. 7, pp. 1–14 (online), DOI: 10.1186/1745-6150-7-12 (2012).

[6] Zimmermann, L., Stephens, A., Nam, S. Z., Rau, D., Kübler, J., Lozajic, M., Gabler, F., Söding, J., Lupas, A. N. and Alva, V.: A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core, *Journal of Molecular Biology*, No. Table 1, pp. 1–7 (online), DOI: 10.1016/j.jmb.2017.12.007 (2017).

[7] Hildebrand, A., Remmert, M., Biegert, A. and Söding, J.: Fast and accurate automatic structure prediction with HHpred, *Proteins: Structure, Function and Bioinformatics*, Vol. 77, No. SUPPL. 9, pp. 128–132 (online), DOI: 10.1002/prot.22499 (2009).

[8] Meier, A. and Söding, J.: Automatic Prediction of Protein 3D Structures by Probabilistic Multi-template Homology Modeling, *PLoS Computational Biology*, Vol. 11, No. 10, pp. 1–20 (online), DOI: 10.1371/journal.pcbi.1004343 (2015).

[9] Kopp, J., Bordoli, L., Battey, J. N., Kiefer, F. and Schwede, T.: Assessment of CASP7 predictions for template-based modeling targets, *Proteins: Structure, Function, and Bioinformatics*, Vol. 69, No. S8, pp. 38–56 (online), DOI: 10.1002/prot.21753 (2007).

[10] Hijikata, A., Yura, K., Noguti, T. and Go, M.: Revisiting gap locations in amino acid sequence alignments and a proposal for a method to improve them by introducing solvent accessibility, *Proteins: Structure, Function, and Bioinformatics*, Vol. 79, No. 6, pp. 1868–1877 (online), DOI: 10.1002/prot.23011 (2011).

[11] Leyi, W. and Quan, Z.: Recent Progress in Machine Learning-Based Methods for Protein Fold Recognition, *International Journal of Molecular Sciences*, Vol. 17, No. 12, p. 2118 (2016).

[12] James, L., Abdollah, D., Rhys, H., Alok, S., Kuldip, P.,

Abdul, S., Yaoqi, Z. and Yuedong, Y.: Predicting backbone C_α angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network, *Journal of Computational Chemistry*, Vol. 35, No. 28, pp. 2040–2046 (2014).

[13] Balachandran, M. and Jooyoung, L.: SVMQA: support-vector-machine-based protein single-model quality assessment., *Bioinformatics*, No. 16 (2017).

[14] Renzhi, C., Debswapna, B., Jie, H. and Jianlin, C.: DeepQA: improving the estimation of single protein model quality with deep belief networks, *BMC Bioinformatics*, Vol. 17, No. 1, p. 495 (2016).

[15] Sheng, W., Siqi, S., Zhen, L., Renyu, Z. and Jinbo, X.: Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model, *PLOS Computational Biology*, Vol. 13, No. 1, p. e1005324 (2017).

[16] Sheng, W., Jian, P., Jianzhu, M. and Jinbo, X.: Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields, *Scientific Reports*, Vol. 6, No. 1, p. srep18962 (2016).

[17] Tomii, K. and Akiyama, Y.: FORTE: A profile-profile comparison tool for protein fold recognition, *Bioinformatics*, Vol. 20, No. 4, pp. 594–595 (online), DOI: 10.1093/bioinformatics/btg474 (2004).

[18] Rychlewski, L., Li, W., Jaroszewski, L. and Godzik, A.: Comparison of sequence profiles. Strategies for structural predictions using sequence information, *Protein Science*, Vol. 9, No. 2, pp. 232–241 (online), DOI: 10.1110/ps.9.2.232 (2000).

[19] Smith, T. F. and Waterman, M. S.: Identification of common molecular subsequences, *Journal of Molecular Biology*, (online), DOI: 10.1016/0022-2836(81)90087-5 (1981).

[20] Murzin, A. G., Brenner, S. E., Hubbard, T. and Chothia, C.: SCOP: A structural classification of proteins database for the investigation of sequences and structures, *Journal of Molecular Biology*, (online), DOI: 10.1016/S0022-2836(05)80134-2 (1995).

[21] Fox, N. K., Brenner, S. E. and Chandonia, J. M.: SCOPe: Structural Classification of Proteins - Extended, integrating SCOP and ASTRAL data and classification of new structures, *Nucleic Acids Research*, Vol. 42, No. D1, pp. 1–6 (online), DOI: 10.1093/nar/gkt1240 (2014).

[22] Zhang, Y. and Skolnick, J.: TM-align: A protein structure alignment algorithm based on the TM-score, *Nucleic Acids Research*, Vol. 33, No. 7, pp. 2302–2309 (online), DOI: 10.1093/nar/gki524 (2005).

[23] Zhang, Y. and Skolnick, J.: Scoring function for automated assessment of protein structure template quality, *Proteins: Structure, Function and Genetics*, Vol. 57, No. 4, pp. 702–710 (online), DOI: 10.1002/prot.20264 (2004).

[24] Xu, J. and Zhang, Y.: How significant is a protein structure similarity with TM-score = 0.5?, *Bioinformatics*, (online), DOI: 10.1093/bioinformatics/btq066 (2010).

[25] Smith, T. and Waterman, M.: Identification of common molecular subsequences, *Journal of Molecular Biology*, Vol. 147, No. 1, pp. 195 – 197 (online), DOI: https://doi.org/10.1016/0022-2836(81)90087-5 (1981).

[26] Wu, X., Kumar, V., Ross, Q. J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z. H., Steinbach, M., Hand, D. J. and Steinberg, D.: Top 10 algorithms in data mining, *Knowledge and Information Systems*, (online), DOI: 10.1007/s10115-007-0114-2 (2008).

[27] Söding, J.: Protein homology detection by HMM-HMM comparison, *Bioinformatics*, Vol. 21, No. 7, pp. 951–960 (online), DOI: 10.1093/bioinformatics/bti125 (2005).

[28] Marchler-Bauer, A., Lu, S., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., Fong, J. H., Geer, L. Y., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Jackson, J. D., Ke, Z., Lanczycki, C. J., Lu, F., Marchler, G. H., Mullokandov, M., Omelchenko, M. V., Robertson, C. L., Song, J. S., Thanki, N., Yamashita, R. A., Zhang, D., Zhang, N., Zheng, C. and Bryant, S. H.: CDD: A Conserved Domain Database for the functional annotation of proteins, *Nucleic Acids Research*, (online), DOI: 10.1093/nar/gkq1189 (2011).

[29] Mirdita, M., Von Den Driesch, L., Galiez, C., Martin, M. J., Söding, J. and Steinegger, M.: Uniclust databases of clustered and deeply annotated protein sequences and alignments, *Nucleic Acids Research*, (online), DOI: 10.1093/nar/gkw1081 (2017).

[30] Šali, A. and Blundell, T. L.: Comparative protein modelling by satisfaction of spatial restraints, *Journal of Molecular Biology*, (online), DOI: 10.1006/jmbi.1993.1626 (1993).