Vol.2017-MPS-113 No.34
Vol.2017-BIO-50 No.34
2017/6/24

配列プロファイル生成の改良によるタンパク質天然変性領域予測の高速化

賀来 智博[1,a)]　石田 貴士[1,b)]

概要：タンパク質の構造予測の 1 つとして、タンパク質で一定の構造をとらない領域の予測を行う天然変性領域予測がある．既存の天然変性領域予測手法は高い予測精度を示しているが、その一方で予測に時間がかかることという点が一つの問題となっている．既存の多くの予測手法では、1 件のタンパク質の予測に数分〜数十分要しており、数千、数万件のタンパク質について予測を行おうとすると数十日を要してしまう．そこで本研究では、タンパク質天然変性領域予測において、予測速度のボトルネックとなっている配列プロファイル作成の高速化を行うことにより、予測精度を落とさずに予測時間の改善を目指す．これを実現するために、配列プロファイル作成に用いる配列データベース内の配列数削減と、配列データベース検索を行わないコンテキスト依存プロファイル作成手法を用いた高速化の 2 点を行った．

Acceleration of protein disordered region prediction by improvement of sequence profile generation

Kaku Tomohiro[1,a)]　Ishida Takashi[1,b)]

Abstract: Disordered region prediction predicts regions that do not take a certain structure in proteins. Although existing methods show high prediction accuracy, they take several minutes to several tens of minutes to predict one protein, and it takes several tens of days to predict thousands or tens or thousands of proteins. In this research, we aim to improve the prediction time without degrading the prediction accuracy by speeding up the sequence profile creation, which is a bottleneck of the predicted speed in disordered region prediction. IFirstly, we reduced the number of sequences in the sequence database used for sequence profile creation. Secondary,we tried to accelerate the predction using a context-specific profile generation method without performing sequence database search.

1. 導入

タンパク質はアミノ酸配列から一意に決まる立体構造を持ち、その立体構造はタンパク質の機能と密接な関わりがある．しかし、近年一定の立体構造をとらないタンパク質、または部分的に立体構造をとらない領域を持つタンパク質が知られるようになった．そのような領域はタンパク質の天然変性領域と呼ばれている．天然変性領域にはタンパク質の機能にとって重要な領域が含まれていることが明らかになっている [1], [2].

タンパク質の天然変性領域を予測することで、タンパク質の重要な機能を知る手がかりになり、また天然変性領域以外の予測可能な配列をあらかじめ知っておくことで、立体構造予測に費やす労力を少なくすることができる．そのため天然変性領域を予測する手法が研究されてきた [3]. 一般的に、天然変性領域予測問題は簡単のために 2 クラスの分類問題として扱われている．タンパク質の天然変性領域のアミノ酸配列には単純な繰り返しや、特定のアミノ酸が通常の領域に比べ多く見られるといった特徴的なパターンが存在していることが過去の研究から明らかになっている．そのため、タンパク質の立体構造予測と同じように、タンパク質のアミノ酸配列から各アミノ酸について天然変性領域を予測することができる [2], [4].

これまで多数の天然変性領域予測手法が提案されてお

[1] 東京工業大学　情報理工学院　情報工学系
[a)] kaku@cb.cs.titech.ac.jp
[b)] ishida@cb.cs.titech.ac.jp

り、それらの予測手法の性能の比較に国際的なベンチマークが行われている．2012 年に行われた最新の天然変性領域予測手法のベンチマークである Critical Assessment of protein Structure Prediction（CASP）10[5] において、良い精度を示している予測手法では、ほとんどがタンパク質のアミノ酸列から作成した配列プロファイルを用いており、それを入力列として機械学習で予測を行っている．配列プロファイルとは各位置におけるアミノ酸の進化的な出現頻度を表したものであり、タンパク質配列からの構造予測に用いることで、予測が顕著に改善されることがわかっている [6], [7]．またこれらの予測では、配列プロファイルを構造予測に用いるとき、配列プロファイルそのものではなく位置特異的スコアマトリクスのかたちで用いられる事が多い．位置特異的スコアマトリクスとは、入力配列について類似配列の探索を行うアルゴリズムである PSI-BLAST[8] によって生成される、入力配列の各列において各アミノ酸残基の現れやすさをスコアとして表したものである．

既存の天然変性領域予測手法では高い予測精度を示している一方で、問題点の一つとして予測に時間がかかることが挙げられる．また、機械学習による予測ではなく、配列プロファイルの生成にほとんどの時間を費やしており、これは全ての予測手法において同様の問題である．

本研究では、配列プロファイルを用いたタンパク質天然変性領域予測において、ボトルネックとなっている配列プロファイル作成の高速化を行うことにより、予測精度を落とさずに予測時間の改善を目指す．また、本研究では天然変性領域予測手法として、CASP10 において良い精度を示した PrDOS-CNF の学習アルゴリズムに、CNF に比べて訓練が容易な SVM を適用した PrDOS[9] を用いて実行時間計測、予測精度の評価を行う．

本論文では第２節で配列プロファイル作成に用いるデータベースの配列数削減による高速化とその結果について述べる．次に、第３節でコンテキスト依存プロファイルを用いた高速化とその結果について述べる．最後に第４節で本研究の結論、今後の課題について述べる．

2. データベースの配列数削減によるタンパク質プロファイル作成の高速化

多くの天然変性領域予測手法では配列プロファイル作成手法として PSI-BLAST を用いている．PSI-BLAST では入力されたタンパク質配列について、配列データベース内で入力配列と配列構造が類似している配列を検索し、検索で見つかった全配列を用いて配列プロファイルを作成する．一般に、検索を行う際の配列データベース内の配列数が多いと、検索にかかる時間が長くなり、配列プロファイル作成に要する時間が長くなる．そこで本研究では配列データベース内の配列数を削減することによって、天然変性領域予測の結果を落とさずに配列プロファイル生成の高速化を図る．本節では、配列数が少ない配列データベースを用いることで高速化を試み、またそれによって予測精度がどのように変化するのか実験を行った．

2.1 PSI-BLAST を用いて作成した配列プロファイル

PrDOS や多くの天然変性領域予測手法では PSI-BLAST を用いて作成した配列プロファイルから位置特異的スコアマトリクスを算出し、それを入力列として SVM で天然変性領域予測を行っている．本節ではまず、PSI-BLAST での配列プロファイル生成の概略を示す．

PSI-BLAST では入力配列に対して、配列データベース内で類似配列の探索を行い、入力配列と得られた類似配列の類似な部分が並ぶように整列を行ったマルチプルアラインメントを作成する．作成したマルチプルアラインメントの各列に現れる各アミノ酸の出現頻度として配列プロファイルが生成される．位置特異的スコアマトリクスは各アミノ酸が出現する確率である背景確率で配列プロファイルを割り、対数をとった対数オッズスコアで算出される．PSI-BLAST では作成した位置特異的スコアマトリクスを用いて再度類似配列の探索を行う．また現在の PSI-BLAST の実装では、配列プロファイルを生成するためには 2 回の配列データベース検索を行なう必要がある [8]．

2.2 用いた配列データベース

PrDOS 内の PSI-BLAST で現在用いられている配列データベースは NCBI-nr である．そこで、本研究では NCBI-nr より配列数が少ない一般的に提供されている配列データベースとして UniProt Reference Clusters（以下 Uniref とする）の Uniref100、Uniref90、Uniref50 と pdbaa の 4 つを用意した．また用意した配列データベースはすべて、配列データベース中に同じアミノ酸配列が含まれていない（非冗長）なものである．各配列データベース中のタンパク質配列数と取得年月を表 1 で示す．NCBI-nr（non redundant）は NCBI が提供している非冗長な配列データベースである [10]．Uniref 系のデータベースは UniProt コンソーシアムが提供している非冗長な配列データベースである．また Uniref90 と Uniref50 は Uniref100 を基準として配列同士の配列一致率 90 %、50 %とそれぞれ設定し、クラスタ代表配列間の一致がその類似度以下となるように CD-HIT[11] を用いてクラスタリングを行った配列データベースである [12]．NCBI-pdbaa は Protein Data Bank に登録されている、立体構造がわかっているタンパク質配列データベースである [13]．

2.3 実験

配列プロファイル作成手法として PSI-BLAST を用いた時に、配列数が少ない配列データベースを用いることで配列プロファイル生成の高速化を試みた．また、それによっ

表 1 配列データベース

配列データベース名	タンパク質配列数	取得年月
NCBI-nr	88M	2016/5
Uniref100	83M	2016/8
Uniref90	43M	2016/6
Uniref50	17M	2016/6
pdbaa	0.08M	2016/7

1M=1,000,000

表 2 TSUBAME2.5 Thin ノード（S キュー）

CPU	Xeonn 5670 2.93GHz 6 cores × 2
Memory	54GB
ローカル SSD	50GB
共有ファイルシステム	Lustre（30TB）

て天然変性領域予測精度がどのように変化するのかを調べるために実験を行った．本実験では天然変性領域予測システムとして PrDOS を利用した．

計算機環境

東京工業大学の計算機システムである TSUBAME2.5 の Thin ノード（S キュー) を使用した．上記環境では、1 ノードおいて最大で 12 コアを用いて計算を行うことができるが、一般的な環境を考慮して本研究では 8 スレッドで実行を計測を行った．また配列データベースは共有ファイルシステム（Lustre) に配置し計測を行った．Lustre とはストレージに関するメタデータを管理する機構と実データを保持する機構を分離することで I/O ののボトルネックの解消を図る共有ファイルシステムである [14]．しかし、Lustre はネットワークを介した共有ファイルシステムであるため、ネットワークアクセスのためディスクへのアクセスに時間がかかり、また複数のユーザーが利用をしているため同じディスクに対して I/O が集中した場合は計測時間にばらつきがでてしまい実行時間計測が正しく行えない．そのため Lustre に配列データベースを配置し行った計測に加えて、配列データベースをより高速な I/O デバイスであるノード内のローカルな SSD に配置した場合の計測を行った．各ディスクの計算環境の詳細を表 2 で示す．また TSUBAME2.5 のローカルな SSD の容量が 50GB であるため、それ以下の容量の配列データベースについてローカルな SSD で実験を行った．

入力配列

CASP10 で用いられた予測ターゲットの一部である T0644 から T0719 の合計 76 タンパク質のうち天然変性領域についての正解ラベルがデータに含まれていなかった 10 タンパク質を除いた、平均配列長が 264 の 66 タンパク質を入力配列とした [15]．

実行時間計測

本研究では各配列データベース、各実行ディスク、各入力配列について 3 回づつ実行を行い、その中央値を結果として採用した．また、それぞれの配列データベースについて、各入力配列の予測を行ったときに得られた結果の合計時間を本実験の結果として示す．各実行について全実行時間と、位置特的スコアマトリクス生成時間の 2 つについて計測を行った．実際の実行では 1 度のタスクで数千数万のタンパク質配列についてまとめて予測を行うため、全タンパク配列の予測時間に対して配列データベースのコピーにかかる時間はとても小さく無視できるので、本実験では配列データベースをローカルな SSD にコピーを行う際の時間は計測時間に含めない．

予測精度評価

本研究では精度評価の指標として、CASP10 で天然変性領域予測手法の精度評価の指標として用いられた ROC 曲線下の面積（Area Under the Receiver Operating Characteristic curve、AUC）を用いた [16]．これは、天然変性領域の比率がそれ以外の領域に比べて非常に少なく、通常の Accuracy では性能の評価が難しいためである．AUC の概要を述べる．天然変性領域予測は天然変性領域であるか、そうでないかの 2 クラス分類問題である．天然変性領域であることを正例、そうでないことを負例とすると分類器が出力する結果は以下の 4 つのパターンが考えられる．

- True Positive（TP）：正例の物を正しく正例だと予測した
- True Negative（TN）：負例の物を正しく負例だと予測した
- False Positive（FP）：正例の物を誤って負例だと予測した
- False Negative（FN）：負例の物を誤って正例だと予測した

分類器がデータセットに対して予測を行ったときの出力結果が、True Positive となった要素の数を#TP、True Negative となった要素の数を#TN、False Positive となった要素の数を#FP、False Negative となった要素の数を#FN とする．ROC 曲線は予測モデルの出力の判定をに用いる閾値を変化させながら、縦軸に True Positive Rate、横軸に False Positive Rate をとった曲線である．True Positive Rate（真陽性率）は正例のものの中で正しく予測できた割合であり、以下の式で求められる

$$\text{True Positive Rate} \equiv \frac{\#\text{TP}}{\#\text{TP} \ + \ \#\text{FN}}$$

False Positive Rate（偽陽性率) とは、負例のものの中で誤って予測をした割合であり、以下の式で求められる．

$$\text{False Positive Rate} \equiv \frac{\#\text{FP}}{\#\text{FP} + \#\text{TN}}$$

ROC 曲線下の面積（Area Under th Curve）は分類器の性能のよさを表している．理想的な分類器、つまり正例と負例を完全に分離できる分類器における ROC 曲線は、原点から (0,1) 上昇し、そこから水平に (1,1) まで続き、AUC は 1.0 となる．また、予測をランダムに行う分類器では ROC 曲線は原点と (1,1) を結ぶ直線となり、AUC は 0.5 となる [16].

2.4 結果と考察

表 3 に各配列データベースとそれぞれを配置したディスクについての位置特異的スコアマトリクス生成時間、全実行時間、AUC についての計測結果を示す．また全実行時間と配列プロファイル生成時間について、NCBI-nr を基準として他の配列データベースを用いた時の高速化倍率を示している．

すべての配列データベースについて配列プロファイル作成時間と全実行時間の差がほぼ等しい．このことから PrDOS における SVM を用いた予測やその他の計算などの、配列プロファイル作成以外にかかる時間が配列データベースにほとんど依存せず、一定であることがわかる．

また Lustre と SSD で実行を行った各配列データベースの配列プロファイル作成時間と全実行時間の差がほとんど等しいことから、高速な I/O デバイスを用いても配列プロファイル作成以外にかかる時間は変化しない事がわかる．

各配列データベースについて Lustre と SSD での配列プロファイル作成時間を比較すると、配列件数が多いほど/scr での配列プロファイル作成時間が減少していることがわかる．これは配列データベース内の配列件数が多いほど配列データベースへのアクセス回数が多いので、読み込みが高速になったときの削減時間も多くなるためと考えられる．今後の研究では研究効率の点から高速な I/O デバイスでの実験を行う．

各配列データベースの配列数と配列プロファイル作成時間から、配列数が少ないほど作成時間が短いことがわかる．しかし、NCBI-nr と Uniref100 の配列プロファイル作成時間を比較すると、配列数が少ない Uniref100 の方が時間を要している．これは計測時間のばらつきによるもの以外に、配列データベース内のタンパク質配列の長さについて、Uniref100 の方が NCBI-nr より長いものが多いからではないかと考えられる．

また NCBI-nr や Uniref100 に対して配列数が半分以下の Uniref90 や Uniref50 が 0.88 以上の AUC を出していることから．配列データベースの配列数と AUC は正の相関関係にないことがわかる．これは、配列データベース中に似たような配列が多いと、特定の配列構造についてスコアが出やすくなり、正確な予測が行えなくなるが、これがクラスタリングによって似たような配列が削減され、解消されたためと考えられる．しかし NCBI-pdbaa の結果からわかるように配列数を減らしすぎると配列プロファイル生成の精度が下がり、予測精度が落ちてしまう．今後どの程度まで AUC を下げずに配列データベースの配列数を削減できるのか、確認を行なう必要があると考えられる．Uniref50 より配列数の少ない配列データベースとしての基準を変更し、Uniref30 や Uniref10 を作成し同様に実行時間、精度の計測を行うことが必要であろう．

今回の実験では各配列データベースの取得年月が異なっており、配列データベース間で正確な比較が出来ていないので、今後同じ時期に配列データベースを取得し再度、評価を行うべきである．しかし、新しい時期に取得した Uniref100 に比べて古い時期に取得した Uniref90 や Uniref50 が良い性能を示していることから、取得時期による配列データベース間の予測精度の優劣は変わらないものと思われる．

3. コンテキスト依存プロファイルを用いた高速化

PSI-BLAST では入力タンパク質配列について配列データベース検索を行い、見つかった相同配列を用いて配列プロファイルを作成する．しかし、配列データベースにおいて相同性検索を行うことは時間を要する．そこで、本節ではあらかじめ様々なタンパク質列に対応できるプロファイルを用意しておき、入力タンパク質配列に対して相同性検索を行わずにコンテキスト依存で配列プロファイルを作成する手法 [17] を用いることで配列プロファイル生成の高速化を図る．本節ではまず、CSI-BLAST によって生成されるコンテキスト依存プロファイルと CSI-BLAST の動作について説明を行う．その後、CSI-BLAST によるコンテキスト依存プロファイルの生成が PSI-BLAST の配列プロファイル生成に比べて高速であることを示す．最後にコンテキスト依存プロファイルを用いて構築した天然変性領域予測モデルと PSI-BLAST で作成した配列プロファイルを用いて構築した予測モデルの精度比較を行う．

3.1 コンテキスト依存プロファイル

コンテキスト依存配列プロファイルとは、2012 年に Angermller らによって提案された CSI-BLAST において、類似配列の探索に用いられている配列プロファイルである．CSI-BLAST は PSI-BLAST に比べて類似配列探索において良い精度を示している [17]．また、このコンテキスト依存配列プロファイルは配列相同性検索を行うこと無く、あらかじめ用意された様々なタンパク質のアミノ酸配列に対応できるプロファイルから生成される．

CSI-BLAST では入力配列に対してまずコンテキスト依存プロファイルを作成する．次に作成したコンテキスト依存プロファイルから PSI-BLAST と同様に位置特異的スコ

表 3 各配列データベースの実行結果

配列データベース	配列数数	実行ディスク	PSI-BLAST（sec）	全実行時間（sec）	AUC
NCBI-nr	88M	Lustre	80,780（× 1.00）	83,072（× 1.00）	0.875
Uniref100	83M	Lustre	81,206（× 0.99）	83,437（× 0.99）	0.877
Uniref90	43M	Lustre	38,985（× 2.07）	41,205（× 2.01）	0.885
		SSD	6,419（× 12.6）	7,662（× 10.8）	
Uniref50	17M	Lustre	13,668（× 5.91）	15,892（× 5.23）	0.888
		SSD	2,107（× 38.3）	3,218（× 25.8）	
NCBI-pdbaa	0.08M	Lustre	55（× 147）	2,279（× 36.5）	0.827
		SSD	27（× 300）	1,118（× 74.3）	

アマトリクスを作成し配列データベースに対して類似配列の探索が行われ、探索に用いられた位置特異的スコアマトリクスが出力される．類似配列の探索前にコンテキスト依存プロファイルが作成されるため、コンテキスト依存プロファイルは PSI-BLAST によって生成される配列プロファイルと異なり類似配列の探索結果によらないものとなる．そのため配列プロファイル生成時間が PSI-BLAST に比べて短くなると考えられる．また CSI-BLAST は類似配列探索において同様のツールである PSI-BLAST より良い性能を示している [17]．以上より、PSI-BLAST に比べて高速に精度を落とさない配列プロファイルを生成する手法として CSI-BLAST を用いて予測精度の評価を行う．

表 4 2iterationPSI-BLAST と 1iterationCSI-BLAST の実行時間

配列プロファイル作成手法	実実行時間（sec）
1 iteration CSI-BLAST	0.28
2 iteration PSI-BLAST	61.84
1 iteration PSI-BLAST（推定）	30.92

3.2 コンテキスト依存プロファイル作成時間計測

CSI-BLAST を用いたコンテキスト依存配列プロファイル作成にかかる時間が、PSI-BLAST によるプロファイル作成時間より高速であるかを調べるために実験を行った．入力配列に対して 2 iteration PSI-BLAST と 1 iteration CSI-BLAST を用いて同じ入力配列に対して配列プロファイルの作成と配列プロファイルから位置特異的スコアマトリクスの生成を行う．iteration とは配列データベース検索を行う回数である．PSI-BLAST について配列プロファイル生成に用いた配列データベースは Uniref90 である．

計算機環境

東京工業大学の計算機システムである TSUBAME2.5 の Thin ノード（S キュー) を使用した．配列データベースをローカルなディスクに配置し 8thread で実行を行った．計算環境の詳細を表 2 で示す．

入力配列

CASP10 の T0644 タンパク質（残基数 166）

配列プロファイル作成に用いた配列データベース

Uniref90 に対して探索を行った．配列データベースの詳細については 2.1 節と表 1 で示している．

実行時間計測

time コマンドで実実行時間の計測 1 回行い、得られた結果をそれぞれの手法の配列プロファイル作成時間として採用する．現在の PSI-BLAST の実装では配列プロファイルを作成するために 2 回の類似配列探索を行う必要がある．しかし、2 回目の類似配列探索は配列プロファイル生成に影響を与えないため、本来は行う必要がない．PSI-BLAST のソースコードを修正することで、2 回目の検索を実行しないようにすることは可能であるが、多大な困難を伴うため、今回の実行時間計測では PSI-BLSAT の実行時間を半分にした値を採用する．

3.3 結果

表 4 に 2 iteration PSI-BLAST と 1 iteration CSI-BLAST に対して time コマンドを用いて解析を行ったそれぞれの実実行時間を示す．PSI-BLAST の実行時間については 2 iteration PSI-BLAST で計測された実行時間の半分を、1 iteration PSI-BLAST の推定実行時間として採用している．

3.3.1 考察

結果より、明らかに 1 iteration CSI-BLAST が 1 iteration PSI-BLAST より高速であることがわかる．よって、天然変性領域予測において 1 iteration CSI-BLAST を配列プロファイル作成手法として用いた時に、2 iteration PSI-BLAST を配列プロファイル作成手法として用いた時に比べて精度が大きく落とさなければ、1 iteration CSI-BLAST を用いたコンテキスト依存プロファイル作成を天然変性領域予測に用いることが有用であると言える．

3.4 予測精度の評価

コンテキスト依存プロファイルを用いた天然変性領域予測を行ったとき、PSI-BLAST で作成した配列プロファイルを用いた天然変性領域予測の精度に比べてどの程度変化するのか調べるために実験を行った．CSI-BLAST が出力するスコアは PSI-BLAST が出力するスコアに比べて値域が広いものとなっている．そのため、PSI-BLAST で訓練

を行った予測モデルに対して CSI-BLAST で作成した位置特異的スコアマトリクスを入力し予測を行うことは適していないと考えられる．

そこで本節では CSI-BLAST でコンテキスト依存配列プロファイルから作成された位置特異的スコアマトリクスを用いて PrDOS と同様に入力データを作成し、SVM の訓練を行うことで予測モデルを生成した．SVM は教師付き学習におけるアルゴリズムの一つである．本研究では実装として LIBSVM(ver3.22) を用いた [19]．SVM の目的関数と rbf カーネルそれぞれについて、ハイパーパラメータを以下のそれぞれの値について設定して、トレーニングセットに対して 10-fold cross validation を行いハイパーパラメータを決定する．各 validation に対して AUROC を算出し、それらの平均を結果として採用する．ここで、C を誤識別率とマージンの比重、γ を rbf カーネルの式

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = exp(-\gamma||\boldsymbol{x}_i - \boldsymbol{x}_j||^2)$$

の中に出てくる変数 γ とする．これら 2 つのハイパーパラメータについて下記の範囲でグリッドサーチで探索を行った．

$$C = \{2^{-5}, 2^{-4.5}, ..., 2^{0.5}, 2^{1}\}$$
$$\gamma = \{2^{-5}, 2^{-4.5}, ..., 2^{0.5}, 2^{1}\}$$

また 3.5 節において示すとおり、天然変性領域とそれ以外の領域のデータセットの比が約 1 対 20 となっているので、天然変性領域のラベルについて 20 の重みをかけて訓練を行った．

3.5 訓練データセット

PISCES[18] を用いて表 5 の条件で PDB に登録されている全タンパク質から 3,021 件のタンパク質構造リストを取得した．取得したタンパク質構造リストについてスクリプトを用いて天然変性領域判定処理 [9] を行い、エラーの出なかった 2,593 件をデータセットとして用いた．

表 5 PISCES を用いた際の設定条件

解像度（resolution）	1.6
R 因子	0.25
取得年月日	2016/10/21
取得配列数	3,021

上記のタンパク質に対して 1 iteration CSI-BLAST を用いて位置特異的スコアマトリクスを作成した．作成した位置特異的スコアマトリクスに対して PrDOS と同様に入力列の各アミノ酸残基に対して、それぞれのアミノ酸残基を中心に前後 13、合計 27 のアミノ酸残基についての位置特異的スコアマトリクスを 1 件の入力データとしてデータセットの作成を行った [9]．

その結果 591,772 件の入力データが得られた．そのうち天然変性領域であるものは 28,021 件であった．このデータセットを全て用いて SVM の訓練を行った場合、1 度の訓練の完了に数日を要するためハイパーパラメータの探索に適していない．そのため全入力データから約 10,000 件前後となるように入力データをランダムに選択し小さいデータセットを作成した．その結果、10,172 件（天然変性領域であるものは 456 件）の入力データを含むデータセットが得られた．この削減したデータセットを用いて SVM のハイパーパラメータを決定し精度の比較を行う．

3.6 配列プロファイル作成手法

用いた配列プロファイル作成手法は 2 iteration PSI-BLAST と 1 iteration CSI-BLAST である．PSI-BLAST について配列プロファイルを作成するのに用いる配列データベースとして、2 節において良い精度を示した Uniref90 と Uniref50 を採用した．

3.7 結果

表 6 に各配列プロファイル作成手法を用いて SVM の訓練を行った結果を示す．結果より、CSI-BLAST を用いて作成したコンテキスト依存プロファイルを用いたときの予測精度の精度は、Uniref90 を配列プロファイル作成に用いた PSI-BLAST に比べて予測精度が低いことがわかる．また、CSI-BLAST を用いて作成したコンテキスト依存プロファイルを用いたときの予測精度は、Uniref50 を配列プロファイル作成に用いた PSI-BLAST に比べてほとんど変わらないことがわかる．

3.8 考察

CSI-BLAST の配列プロファイル生成時間は PSI-BLAST に比べて早く、精度も十分である．そのため実際の運用では、精度重視の予測を行うか、速度重視の予測を行うかで使い分ける事が可能であると思われる．

4. まとめ

4.1 配列データベースの削減によるタンパク質プロファイル作成手法の高速化

配列プロファイル作成手法として PSI-BLAST を用いたとき、クラスタリングによって Uniref100 から配列数を削減することで作成した Uniref90、Uniref50 を用いると天然変性領域予測の精度を落とすこと無く高速化を行うことが可能であることが示された．また、高速な I/O デバイスを用いることで配列データベースの読み込みを高速化し、配列プロファイル作成時間を削減することができることもわかった．

表 6　各プロファイル生成手法での予測精度

配列プロファイル作成手法	配列プロファイル作成に用いた配列データベース	C	γ	AUROC
PSI-BLAST	Uniref90	$2^{-2.5}$	$2^{-4.5}$	0.9007
	Uniref50	$2^{-0.5}$	2^{-4}	0.8867
CSI-BLAST	なし	$2^{-1.5}$	$2^{-4.5}$	0.8852

4.2 コンテキスト依存配列プロファイル作成手法を用いた高速化

CSI-BLAST を用いたコンテキスト依存プロファイルは PSI-BLAST を用いた配列プロファイル生成より高速であることが示された．また、CSI-BLAST を用いた天然変性領域予測の精度は PSI-BLAST を用いたときと比較して、大きく精度を落とさないことが示された．

4.3 今後の課題

第 2 節の配列プロファイル作成手法として PSI-BLAST を用いたときの実験の結果から、どの程度配列データベースの配列数をクラスタリングで削減しても精度が落ちないのか調べるために、Uniref50 より小さい配列データベースとして Uniref30 と Uniref10 をクラスタリングを用いて作成し時間と精度の計測を行う．

また、第 2 節の実験で用いた配列データベースの取得時期が異なっているので、精度について配列データベース間で正確な比較ができていないため、同じ時期に配列データベースの取得を行い時間と精度の計測を行う．

第 3 節において PSI-BLAST と CSI-BLAST を用いて一部のデータセットで SVM の訓練を行ったが、全データセットを用いて訓練を行った場合、各配列プロファイル作成手法について予測精度にどの程度差が出るのか調べるために、全データセットを用いて各配列プロファイル作成手法について SVM の訓練を行い、CASP10 のタンパク質を入力配列として予測を行い、実行時間と精度を計測する．

参考文献

[1] Ward, Jonathan J., et al. ”Prediction and functional analysis of native disorder in proteins from the three kingdoms of life.” Journal of molecular biology 337.3 (2004): 635-645.
[2] Romero, Pedro, et al. ”Sequence complexity of disordered protein.” Proteins: Structure, Function, and Bioinformatics 42.1 (2001): 38-48.
[3] Linding, Rune, et al. ”Protein disorder prediction: implications for structural proteomics.” Structure 11.11 (2003): 1453-1459.
[4] Uversky, Vladimir N. ”Natively unfolded proteins: a point where biology waits for physics.” Protein science 11.4 (2002): 739-756.
[5] Kryshtafovych, Andriy, et al. ”Assessment of the assessment: evaluation of the model quality estimates in CASP10.” Proteins: Structure, Function, and Bioinformatics 82.S2 (2014): 112-126.
[6] Webb, Benjamin, and Andrej Sali. ”Protein structure modeling with MODELLER.” Protein Structure Prediction (2014): 1-15.
[7] Rost, Burkhard, and Chris Sander. ”Prediction of protein secondary structure at better than 70% accuracy.” Journal of molecular biology 232.2 (1993): 584-599.
[8] Altschul, Stephen F., et al. ”Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.” Nucleic acids research 25.17 (1997): 3389-3402.
[9] Ishida, Takashi, and Kengo Kinoshita. ”PrDOS: prediction of disordered protein regions from amino acid sequence.” Nucleic acids research 35.suppl 2 (2007): W460-W464.
[10] Pruitt, Kim D., Tatiana Tatusova, and Donna R. Maglott. ”NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.” Nucleic acids research 35.suppl 1 (2007): D61-D65.
[11] Li, Weizhong, and Adam Godzik. ”Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.” Bioinformatics 22.13 (2006): 1658-1659.
[12] Suzek, Baris E., et al. ”UniRef: comprehensive and non-redundant UniProt reference clusters.” Bioinformatics 23.10 (2007): 1282-1288.
[13] Berman, Helen M., et al. ”The protein data bank.” Nucleic acids research 28.1 (2000): 235-242.
[14] http://lustre.org（参照 2017–01–28）
[15] Kryshtafovych, Andriy, Bohdan Monastyrskyy, and Krzysztof Fidelis. ”CASP prediction center infrastructure and evaluation measures in CASP10 and CASP ROLL.” Proteins: Structure, Function, and Bioinformatics 82.S2 (2014): 7-13.
[16] Lobo, Jorge M., Alberto Jimnez - Valverde, and Raimundo Real. ”AUC: a misleading measure of the performance of predictive distribution models.” Global ecology and Biogeography 17.2 (2008): 145-151.
[17] Angermller, Christof, Andreas Biegert, and Johannes Sding. ”Discriminative modelling of context-specific amino acid substitution probabilities.” Bioinformatics 28.24 (2012): 3240-3247.
[18] Wang, Guoli, and Roland L. Dunbrack. ”PISCES: a protein sequence culling server.” Bioinformatics 19.12 (2003): 1589-1591.
[19] Hsu, Chih-Wei, Chih-Chung Chang, and Chih-Jen Lin. ”A practical guide to support vector classification.” (2003): 1-16.