研究論文

天然変性領域予測におけるプロリンの重要性†

家富花奈*，安保勲人**，伊藤駿介*，福地佐斗志*,**

The importance of proline residues for the prediction of intrinsically disordered regions†

Kana Ietomi*, Hiroto Anbo**, Shunsuke Ito*,

and Satoshi Fukuchi*, **

Intrinsically disordered regions (IDRs) can be predicted by computer programs. In this work, we pursued what factors provide basis for predicting IDRs. We conducted a random forest analysis to obtain degrees of contribution of each of the amino acid residues for the predictions. The results suggested that the contribution of proline is remarkably larger than other residues. Next, we analyzed the distribution of proline residues around the boundaries between IDRs and structural domains (SDs), disclosing that proline residues notably overrepresent in the SD sides of the boundaries. This result can contribute to develop more accurate prediction programs and to understand the structural nature of intrinsically disordered proteins.

Key words : Bioinformatics, Intrinsically Disordered Protein, Machine Learning

1　はじめに

タンパク質はアミノ酸が重合してできた分子であり，この数珠が立体的に折り畳まれ立体構造を形成し機能を発揮する．タンパク質に関するこのような描像は普遍的だと思われていたが，今世紀に入り描像から逸脱したタンパク質が知られるようになった．天然変性タンパク質(intrinsically disordered protein）は，生理的環境下にあっても単独では立体構造を形成しない．それでいて天然変性タンパク質はちゃんと機能を発揮する[1,2]. 天然変性タンパク質にはアミノ酸配列の全域にわたり立体構造を形成しないものもあるが，一般的には数十から数百残基にも及ぶ柔らかな天然変性領域と立体構造を持つ部分構造ドメインの組み合わせからなっている．天然変性領域はアミノ酸組成に特徴があり，計算機プログラムを使って予測できる．前橋工科大学福地研究室では，天然変性領域予測プログラム NeProc を開発中であり，ベンチマークプログラムである Disorepd3[3]を凌駕する性能を発揮している．NeProc は機械学習手法であるニューラルネットワークを用い予測モデルを作成している．本研究では，NeProc がどのようにして天然変性領域と構造領域を見分けているかを解析することで，天然変性タンパク質の構造的特徴や機能を本質的に理解することを目指す．

2　方法

2・1　入力データ

NeProc はホモロジー検索プログラム PSI-Blast[4]内で利用される PSSM(position specific score matrix)を入力として利用している．PSSM は重み行列の一種で，入力であるアミノ酸配列の各サイトについて，blast によるホモロジー検索で得られたヒット配列のアミノ酸頻度をまとめたものである．すなわち，PSSM は配列長×20 次元のベクトルとなる．PSSM により，アミノ酸配列の各サイトについて許容されるアミノ酸の種類や進化的保存度などの情報が得られ，この情報をもとに NeProc は天然変性領域と構造領域を区別している．NeProc の予測モデル構築には各アミノ酸残基について天然変性領域であるか，構造領域であるか，の情報が必要である．一般にこのような「正解」情報は「ラベル」と称される．NeProc では立体構造データベース PDB[5]および天然変性タンパク質データベース Disport[6]から抽出した 2400 個のタンパク質，総残基数 350 万の PSSM を利用している．PDB については，X 線結晶構造解析で座標を決定できなかった残基を天然変性領域とし，その他の領域を構造領域としてラベルをつけた．Disprot 由来データでは，Disprot のアノテーションに従いラベルを付加した．

† 原稿受理 令和２年２月２８日　Received　February　28，2020

* 生命情報学科（Department of Life Science and Informatics）

** 生命情報学専攻（Graduate school of Engineering, Division of Life Science and Informatics）

2・2　ランダムフォレストによる判別

ランダムフォレストは機械学習法の一つであり，ニューラルネットワーク同様に判別問題に適用可能である．ランダムフォレストは，入力データから多くの決定木をランダムに作成し，これらの結果を統合して最終的な結果を出す．二群の判別を行う場合，決定木はサンプルデータを順次分割しクラスター化してゆく．この際，クラスタに分割する基準は，各クラスターがなるべくどちらかの群に偏るように，つまり群の純度が高くなるように分割して行く．決定木は視覚的に分割を確認できるといった利点があるが，外れ値に弱いという欠点がある．しかし，ランダムフォレストはランダムにサンプリングを行うことで，この外れ値の影響を少なくしている点が優れている．また，ランダムフォレストでは各特徴量の寄与率を算出することができ，この点で，今回のように天然変性領域予測のキーとなる特徴量を知るのに最適である． ランダムフォレストの解析は,プログラミング言語 python の scikit-learn ライブラリ [7]を用い実装した．

2・1 に示した入力データを用いランダムフォレストの予測モデルを構築した．その後，テストデータとして天然変性タンパク質データベース IDEAL[8]および立体構造予測コンテスト CASP10[9]の天然変性領域予測の問題を入力し，判別具合をテストした．

3　結果

3・1 ランダムフォレストによる判別

Table1 にランダムフォレストと NeProc の天然変性領域予測の性能評価を示す．評価はマシューの相関係数(Matthew's correlation coefficient，以下 MCC)で示した．

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (1)$$

MCC は(1)式で表され，TP は真陽性の数，TN は真陰性の数，FP は偽陽性の数，FN は偽陰性の数であり，MCC が 1 に近いほど性能が良いことを示している．ランダムフォレストによる判別は, NeProc には及ばないものの，IDEAL データに対しては実用レベルの性能が出ている．このため，このランダムフォレストモデルも一定の予測性能を持つと考え，次に各残基の寄与率を見積もった．

Table 1. Performances of the IDR predictions.

	ランダムフォレスト	NeProc
IDEAL	0.476	0.576
CASP10	0.286	0.315

3・2　各残基の予測結果への寄与率

各残基の予測結果への寄与率を Fig.1 に示す．寄与率は構築した決定木について，ある特徴量をランダマイズした場合，もとの決定木からどれほど性能の劣化が見られるかを評価している．つまり，性能劣化の大きい特徴量は判別に重要と考えられる．Fig.1 を見るとプロリン(P)セリン(S)グルタミン酸(E)リジン(K)等の寄与率が高く，特にプロリンが突出して大きいことがわかる．

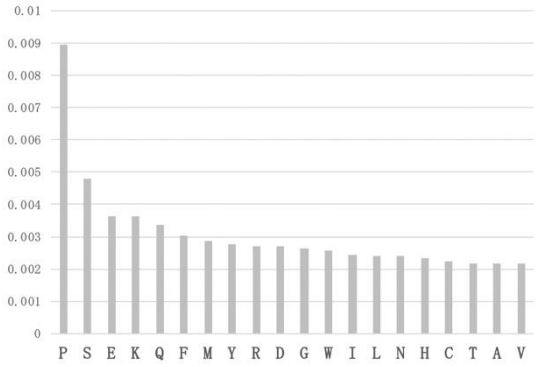

Fig. 1. Contributions of each of the residues in the random forest prediction.

4 考察

4・1 組成の比較

一般にタンパク質には 20 種類のアミノ酸が均等に含まれているわけではなく，また，天然変性領域・構造領域で組成が異なることも知られている．そこで，Fig.2 に天然変性領域と構造領域の組成の差を示す．この図で正の値を持つ残基は天然変性領域に多いことを表す．差の大きいものからセリン，グルタミン酸，プロリン，リジンと並んでおり，Fig.1 で見られた寄与率の大きい残基と傾向が一致している．入力データである PSSM は，各サイトでのアミノ酸残基の出現頻度を数値化したものであることを考えると，このように両者で存在量に差のある残基が判別に重要であることは納得できる．しかし，

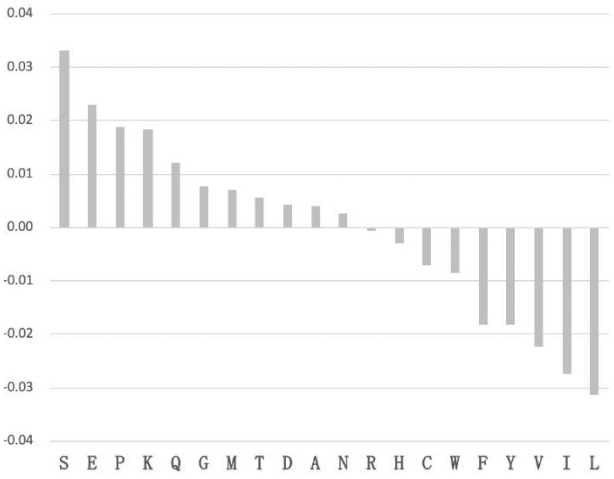

Fig. 2. The differences of the amino acid composition between IDRs and SDs.

Fig.1 に見られるプロリン残基の寄与の大きさは，さらなる考察を要する．そこで，特にプロリンに絞り以下の考察を行った．

4・2 天然変性領域と構造領域の境界の組成

PSSM データを吟味してみると，天然変性領域と構造領域の境界にプロリンが位置している例が多く見られた．そこで，この観察がデータ全体に見られる傾向か否かを判断するため境界付近のアミノ酸頻度を解析した．この解析では，境界から 5, 7, 10 残基(window)の中の各アミノ酸の頻度を以下の式(2)で求めた．ここで，N_A^w は window 中のアミノ酸 A の数，N_A は配列全体のアミノ酸 A の数，N_w は window に含まれる全アミノ酸の数である．境界の片側は天然変性領域，逆側は構造領域となるので，これらを別々に解析した．また，比較対象としてアミノ酸配列をランダムに混合した仮想配列を生成し，境界の位置はそのままで上記同様に(2)式で頻度を求めた．この仮想配列と現実の配列を比較することで，各アミノ酸が偶然からどれくらいずれて境界に分布しているかを評価できる．

$$\frac{N_A^w}{N_A} \times \frac{1}{N_w} \qquad (2)$$

Fig.3 に結果を示す. Fig.3 では式(2)で求めた各残基の頻度を実際の配列と仮想配列間で差を取りプロットしている．正となる残基は，ランダム配列より実際の配列で多く見られることを表す．天然変性領域のグラフ(Fig.3 a, c, e)では実際の配列より顕著に多く見られる残基が認められるものの，window5, 7, 10 と境界からの距離を伸ばしていっても，その傾向はほとんど変化しない．これに対し構造領域(Fig.3 b, d, f)では，仮想配列との差が天然変性領域より小さいこと，window の違いで頻度差に変化があることが観察される．中でもプロリンの頻度変化は顕著で window5 で他の残基と比べ際立って大きな値を示し，この値は window を大きくするに従い小さくなる傾向にある(Fig.3 b, d, f 中，特に黒で表示)．このことは，プロリンは天然変性領域と構造領域の境界の構造領域側に，ランダム配列に比べ顕著に多く分布していることを示している．

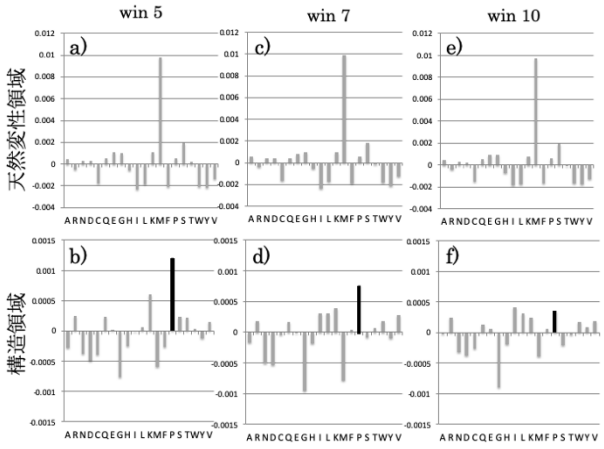

Fig. 3. The frequencies of each of the residues around the boundaries.

また，天然変性領域では全ての window に見られる傾向として，天然変性領域に多く見られるアミノ酸の頻度が大きくなっている．これは，天然変性領域ではアミノ酸配列上での位置にかかわらず，アミノ酸組成は均一であることを示唆している．一方，構造領域で window ごとに各アミノ酸残基の頻度が異なることは，構造領域ではアミノ酸配列上の位置により組成が異なることを示唆しており，立体構造を形成するという機能的制約が影響していることが考えられる．

5 まとめ

本研究から，PSSM を入力として機械学習を用いて天然変性領域を予測する際に，天然変性領域と構造領域の全体的なアミノ酸組成に加え，境界部のアミノ酸組成も判別のキーとなっていることが示唆された．特に，構造領域の境界付近にはプロリンが多く分布し，この組成の偏りが予測に大きく寄与している可能性がある．今回の解析はランダムフォレストの予測モデルを元に解析したが，NeProc は実際にはニューラルネットワークを用い予測を行なっている．Table 1 に見られる予測精度の差をみると，予測プログラムとしてはニューラルネットワークを用いるのが現実的であり，ニューラルネットワークモデルについて，同様の解析を行うことが理想的である.ニューラルネットワーク特にディープラーニングは，近年，画像認識に関連する発展が著しい．実際，画像認識分野ではネットワークの内部を解析しモデルの意味を解析することが試みられており [10]，今後，本研究への応用も検討する価値がある．このような課題があるとはいえ，本研究で発見した境界領域，特に構造領域の境界近辺でのプロリンの多さは事実であり，今後天然変性タンパク質の構造理解に役立つ可能性が高い．

謝辞

本研究の一部は，前橋工科大学リサーチ・アシスタント制度の支援を得て行われた．

参考文献

1) K. A. Dunker, et al., Intrinsic disorder and protein function, Biochemistry, **41(21)**, 6573-6582, (2002).
2) P. E. Wright and J. H. Dyson, Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J. Mol. Biol. **293(2)**, 321 – 331, (1999).
3) D. T. Jones and D. Cozzetto, DISOPRED3: precise disordered region predictions with annotated protein-binding activity. Bioinformatics, **31(6)**, 857 – 863 (2015).
4) S. F. Altschul, et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res., **25(17)**, 3389 – 3402 (1997).
5) wwPDB consortium, Protein Data Bank: the single global archive for 3D macromolecular structure data,

Nucleic Acids Res. **47(D1)**, D520 – D528 (2019).

6) A. Hatos et al., Nucleic Acids Res., DisProt: intrinsically disorder annotation in 2020, Nucleic Acids Res., **48(D1)**, D269 – D276 (2019).
7) F. Pedregosa, et al., Scikit-learn: Machine learning in Python, JMLR, **12(Oct)**, 2825 – 2830 (2011).
8) S. Fukuchi et al., IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners, Nucleic Acids Res.**42(D1)**, D320 – D325 (2014).
9) B. Monastyrskyy et al., Assessment of protein disorder regions predictions in CASP10, Proteins, **82(Suppl 2)**, 127 – 137 (2014).
10) S. Carter et al., Exploring neural networks with activation atlases, Distill, **6(Mar)**, (2019) DOI:10.23915/distill.00015.