様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　２７　日現在

機関番号：　３２６８４
研究種目：　基盤研究(C)（一般）
研究期間：　2013～2016
課題番号：　２５３３０３５２

研究課題名（和文）タンパク質ディスオーダー領域における機能部位予測法の研究開発

研究課題名（英文）Development of a method for identifying functional site in disordered regions of proteins

研究代表者

野口　保（NOGUCHI, Tamotsu）

明治薬科大学・薬学部・教授

研究者番号：００３５７７４０

交付決定額（研究期間全体）：（直接経費）　　3,800,000円

研究成果の概要（和文）：タンパク質ディスオーダー領域予測法(POODLE-SとPOODLE-L)および、タンパク質二次構造予測法（PSIPRED)、タンパク質可溶性予測法（ESPRESSO）、タンパク質機能DB(Pfam)を用いたタンパク質ディスオーダー領域における機能部位予測法を開発した。現状の予測精度（Accuracy）は78%を得たが、True Positive Rate(TPR)は43%であった。目標の予測精度（70%）を達成することはできたが、既存の手法のTPR（57%）に比べて低い。
また、本手法を一般公開するためのウェブシステムを構築した。（ただし、現状では予測性能が不十分のため、未公開中）

研究成果の概要（英文）：We developed a method for identifying functional site in disordered regions of proteins by using POODLE-S, POODLE-L, PSIPRED, ESPRESSO and Pfam database. POODLE-S and PODDLE-L are methods for predicting short and long disordered regions of proteins, respectively. PSIPRED is a method for predicting protein secondary structures developed by David Jones, et al. ESPRESSO is a system for estimating protein expression and solubility in protein expression systems. Pfam database, is available at EMBL-EBI site, is a large collection of protein families, each represented by multiple sequence alignments and identified their function sites. The prediction accuracy of this method is 78% that is better than desired value (70%), but the True Positive Rate (TPR) of this method is 43% that is less than TPR of other method (57%).
The web system to implement the method has developed in our server. But the system is not available, since the prediction efficiency is not sufficient to use.

研究分野：　バイオインフォマティクス

キーワード：　蛋白質　機械学習　生体生命情報学　分子認識　天然変性

１．研究開始当初の背景

研究開始当初、タンパク質の天然変性領域(ディスオーダー領域とも呼ばれる)が、様々な機能に関与する重要な領域であることが実験的に明らかになってきていた。

一方、Ward らは、彼らのディスオーダー領域予測法により、高等生物にこのような領域が特に多く見られる傾向があり、転写調節に関するタンパク質や DNA 結合タンパク質などに多く存在することを示唆し、他の予測法も同様の結果を示した。さらに、Minezaki らは、ヒトのゲノムに対してディスオーダー領域に関する詳細な解析を行い、ヒトの転写因子が原核生物のそれよりディスオーダー領域を多く含むことを示した。また、Hayne らは、真核生物において、ディスオーダー領域を持つタンパク質が、タンパク質相互作用ネットワークのハブになっていることを示した。

上記のように、ディスオーダー領域を持つタンパク質の中には、機能発現の際に立体構造を形成するものがある。そのようなタンパク質は、様々な分子と相互作用できるように、機能部位がディスオーダー領域にあると考えられる。Dunker グループの Mohan らは、ディスオーダー領域の機能部位の構造形成パターンがタンパク質によって異なり、その機能と結合時の構造によって特徴分類できることを示した。また、Vacic らは、その機能部位と結合する相手のタンパク質の特徴を解析し、構造既知の結合部位の形状を基に、タンパク質‐タンパク質結合予測を試行していた。

タンパク質ディスオーダー領域における機能部位予測は、確立されていないが、可能になれば、その部位のアミノ酸配列を基に、タンパク質の機能阻害物質の開発などに有用と考えられる。

我々は、ディスオーダー領域が、配列全体、その長短で、その性質が異なることに着目し、各々のディスオーダーに対応した予測法（POODLE-W,-L,-S：W は配列全体がディスオーダー領域、L は長いディスオーダー領域、S は短いディスオーダー領域）を開発し、それぞれ既存の手法の精度を上回る結果を得た。ディスオーダー領域の性質の差についての考え方は、Mohan らの結果と一致している。

一方、タンパク質の機能発現（他分子との結合）の際に構造変化を伴うことに着目し、構造変化を行う部位を予測する方法を開発し、その結果から機能部位を予測する試みを行った。

２．研究の目的

通常状態で、タンパク質の一部分で立体構造を形成しないディスオーダー領域が、様々な機能に関与する重要な領域であることが実験的に明らかになってきた。それらは、様々な分子と相互作用できるように、通常状態では変性していると考えられている。この領域をアミノ酸配列から予測できれば、その部位を創薬ターゲットにした研究を進展させることができる。

本研究では、我々が開発したディスオーダー領域予測法（POODLE）を基に、「タンパク質ディスオーダー領域における機能部位予測法」を開発し、その予測精度を実用可化レベルにすることを目的とする。

３．研究の方法

(1)　データセット

本研究では、Dunker グループの Fatemeh らによる先行研究（MoRFpred）で用いられたディスオーダー領域内の機能部位データセット（TEST419:419 chains と TEST2012: 45 chains）を用いて性能を比較した。

(2)　手法

まず、タンパク質ディスオーダー領域予測法（POODLE-S, POODLE-L）を用い、ディスオーダー領域内の機能部位予測を行い、その精度を評価する。さらに、タンパク質可溶性予測（ESPRESSO）とタンパク質二次構造予測（PSIPRED）とタンパク質機能部位のデータベース（Pfam）を組み合わせることによって、機能部位予測を行い、その精度を評価する。

機能部位の予測は、以下の条件に合った残基を機能部位とした。

① POODLE-L のディスオーダー予測領域で、POODLE-S のディスオーダー予測領域でない領域にある
② ESPRESSO の可溶性予測において、可溶性に Positive と予測された領域にあることを機能部位の条件に加える
③ PSIPRED の予測でαヘリックスと予測された領域を機能部位の予測から除く
④ Pfam の機能部位領域であることを機能部位の条件に加える

４．研究成果

タンパク質ディスオーダー領域予測法(POODLE-S と POODLE-L)および、タンパク質可溶性予測法（ESPRESSO)、タンパク質二次構造予測法（PSIPRED)、タンパク質機能 DB(Pfam)を用いたタンパク質ディスオーダー領域における機能部位予測法を開発した。

可溶性予測おける Positive の導入、二次構造予測におけるαヘリックスの除外、機能データベースの機能部位の導入は、予測精度向上に効果がなく、本報告では、機能部位の条件①「POODLE-L のディスオーダー予測領域で、POODLE-S のディスオーダー予測領域でない領域にある」のみで予測を行った結果を示す。

データセット TEST419 では、419 chains 中、389 chains の予測結果を得た。TEST2012 45 chains 中 43 chains の予測結果を得た。予測結果が得られなかったのは、そのタンパク質 chain の PSI-BLAST の検索で、検索されたタンパク質件数が多いため、その後の処理が停止し、POODLE-S と PSIPRED の予測結果が得られなかったことが原因であった。

両データベースのテストデータを用いた予測精度（Accuracy）は 78％、True Positive Rate(TPR)は 43%であった。目標の予測精度（Accuracy）の 70%以上を達成することはできたが、既存の手法の TPR は、57%で、それに比べて低い性能しか得られなかった。

それぞれのテストセットで、既存の手法と比較した結果を表１と表２に示す。ACC、TRP、FRP は以下の計算式で求めた。

$$TPR = \frac{TP}{TP+FN}, \qquad FPR = \frac{FP}{TN+F}$$

$$ACC = \frac{1}{2}(TPR + 1 - FRP)$$

ここで、TP、TN、FP、FN は、それぞれ True Positive、True Negative、False Positive、False Negative の数である。

表 1 タンパク質ディスオーダー領域における機能部位予測法の性能（テストセット：TEST419）
*) 389/419 chains の予測結果

Method	*ACC*	*TPR*	*FPR*
MPSPSSMPred	*0.636*	*0.491*	*0.219*
MoRFPred	*0.603*	*0.254*	*0.049*
ANCHOR	*0.568*	*0.389*	*0.253*
This method[*)]	*0.610*	*0.434*	*0.215*

表 2 タンパク質ディスオーダー領域における機能部位予測法の性能（テストセット：TEST2012）
*) 43/45 chains の予測結果

Method	*ACC*	*TPR*	*FPR*
MPSPSSMPred	*0.702*	*0.575*	*0.172*
MoRFPred	*0.596*	*0.236*	*0.045*
ANCHOR	*0.599*	*0.433*	*0.236*
This method[*)]	*0.610*	*0.443*	*0.224*

本研究と並行して行った C. Fang による MFSPSSMpred の ACC、TPR の値が共に最も高い高く、良い性能を示している。本研究で開発した手法は、ACC、TPR の値ではそれ以外の手法を上回ることができた。それに対し、FPR では、MoRFPred が、非常に良い性能を示している。今後、TPR の値を高く維持しつつ、FPR の値を下げることが必要である。

また、本手法を一般公開するためのウェブシステムを構築し、現状の予測手法を実装した。図１は、タンパク質ディスオーダー領域における機能部位予測法公開ウェブシステムの予測起動画面である。現状では用いていないが、タンパク質可溶性予測のオプションを指定して、予測タンパク質のアミノ酸配列を入力して起動ボタンをクリックすると予測計算を開始する。

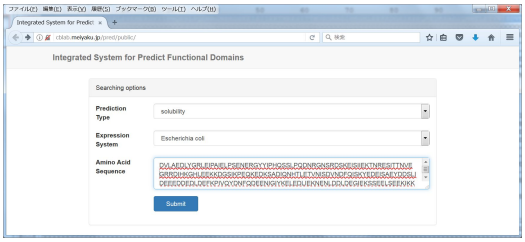

図 1 タンパク質ディスオーダー領域における機能部位予測法公開ウェブシステム起動画面

予測結果の表示画面を図 2 に示す。POODLE-S、POODLE-L、ESPRESSO、PSIPRED を示し、機能部位予測結果を示した後に Pfam の機能部位の情報を表示する。

現状では予測性能が不十分であり、計算の途中で止まってしまう場合があるため、未公開中である。

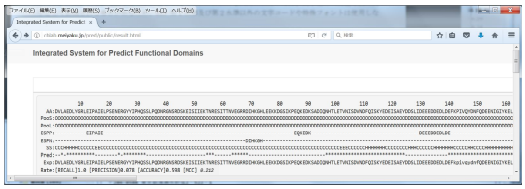

図 2 タンパク質ディスオーダー領域における機能部位予測法公開ウェブシステム結果表示画面

問題が解決し次第、一般に公開する予定である。

５．主な発表論文等

〔雑誌論文〕（計 0 件）

〔学会発表〕（計 0 件）

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計 0 件）

〔その他〕
ホームページ等
http://cblab.meiyaku.jp/CAPPUCCI(予定)

６．研究組織

(1)研究代表者

野口　保　　（NOGUCHI, Tamotsu）
明治薬科大学・薬学部・教授
研究者番号：00357740

(2) 連携研究者

廣瀬　修一　（HIROSE Shuichi）
長瀬産業株式会社
研究者番号：60549898