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　１６　日現在

機関番号：　３４３１５
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2016
課題番号：　２６３３０３３６

研究課題名（和文）機械学習によるタンパク質翻訳後修飾の予測と天然変性領域の機能の解明

研究課題名（英文）Prediction of phosphorylation sites in human protein by machine learning and the functional role of intrinsically disordered regions

研究代表者

西川　郁子（NISHIKAWA, IKUKO）

立命館大学・情報理工学部・教授

研究者番号：９０２１２１１７

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：ヒトタンパク質のリン酸化に対して、修飾部位の機械学習による予測を通して、天然変性領域(IDR)とドメインでの差に着目し，特に天然変性領域がもつ機能的役割を検討した。進化的保存性が低い領域ながら重要な翻訳後修飾も担うIDRは、部位ごとに保存性が異なり、特に機能性が確認されている修飾部位では保存性が高いことが確認できた。領域保存性を前提にできないIDRでも合理性をもつ部位特異的保存度として、複数種のオーソログを用いた定義を提案し、定量的解析を実現した。同時に、リン酸化部位も機能性の有無で分類すると、保存度に差が見られた。保存度情報も用いた機能性リン酸化部位予測は、IDRで精度82.1%を実現した。

研究成果の概要（英文）：Phosphorylation site in human protein is studied, through the prediction by support vector machine. We focus on the difference in evolutional conservation between intrinsically disordered region (IDR) and the domain, and the functional role of IDR in the post-translational modification.
Sequence conservation is known to be generally low in IDR, while the functionally important modifications are often found in IDR. We proposed a measure of site-specific conservation based on multiple ortholog proteins, as PSSM (Position Specific Scoring Matrix), which is often used as a site-specific conservation measure, assumes sequence conservation which does not work in IDR. Then, the site-specific conservation is found to vary within IDR. The conservation is kept high at a phosphorylation site, especially at the phosphorylation with any clarified function. Prediction accuracy improves to 82.1% using both the conservation and the sequence information in IDR for functional phosphorylation sites.

研究分野：　知能情報学

キーワード：　機械学習　ヒトタンパク質　天然変性領域　リン酸化　サポートベクターマシン　進化的保存度　オーソログ　予測

１．研究開始当初の背景

(1) タンパク質への修飾とその機能の解明は、ゲノムサイエンスの最終目標である生命システムの理解への主要課題のひとつである。代表的な翻訳後修飾であるリン酸化や糖鎖修飾では、リン酸や糖鎖がタンパク質に結合することで、多様な生命現象における重要な役割と機能を果たしている。本研究開始までに、*O* 型糖鎖修飾に着目し修飾部位の予測を行った。糖鎖修飾のなかでも *O* 型は、関与する酵素の多様性ゆえに生化学実験のみでは詳細な修飾機構が未解明であることから、機械学習による予測を実施した。

(2) 修飾部位の密度に基づく新たな分類をもとに機械学習による修飾部位予測を行いその有効性を示した。同時に、機能性が注目を集めていたタンパク質の intrinsically disordered 領域(IDR)との関係を初めて定量的に求め、構造安定化・機能多様性との関連を示した。この手法を、最も重要な翻訳後修飾であるリン酸化に適用することで、修飾機構と天然変性領域がもつ機能的役割の解明を目指した。

２．研究の目的

(1) ヒトタンパク質のリン酸化に代表される翻訳後修飾に対して、修飾部位の機械学習による予測を通して、天然変性領域(IDR)とドメインでの差異に着目し、特に天然変性領域が修飾機構に果たす機能的役割の解明を目指す。

(2) ドメインにおける修飾では、コンセンサス配列など配列の規則性で修飾部位が特徴付けられることが予想されるのに対して、IDR における修飾では、領域全体の進化的保存性の低さから、保存度が鍵となることが予想される。定量的評価指標としても、修飾部位の予測精度を用いて、IDR の機能性と保存性の理解につなげる。

３．研究の方法

(1) ヒトタンパク質のリン酸化を対象にドメインと IDR に分けて予測を行う。一つのタンパク質配列をドメインと IDR に分ける指標やアルゴリズムは複数存在するが、研究協力者の西川建博士や福地佐斗志博士ら(Fukuchi S. et al, 2011)は、それらを統合して全タンパク質のアミノ酸残基を二分類するアルゴリズムを開発し、ソフトウエアとして実装・公開（DICHOT）するとともに、ヒトタンパク質全配列をドメインと IDR に二分したデータベースを構築した。それを用いることで、初めて両者を分けた予測が実施可能となった。

(2) 配列保存性が低い領域における部位特異的保存度を新たに定義し用いる。ヒトからの進化的な距離を考慮した複数種の脊椎動物のオルソログを揃え、ヒトタンパク質に対してマルチプル・アラインメントを行い、オルソログ上で対応する各部位でのアミノ酸出現頻度によって保存度を決めた。保存度として一般に用いられる position-specific scoring matrix (PSSM)は、従来ドメインを対象としてきたため、配列の保存を前提として PSI-BLAST から算出される。もともと配列保存性が低い IDR では、PSSM の意味が明確でないと考えるからである。

(3) リン酸化が実験的に見出された部位も、機能が明確なものと、現段階で不明のものに二分する。近年のシーケンシング技術により、非常に多数のリン酸化部位が報告されているが、その中で機能が特定できたものはごく少ない。これらは未だ実験的に特定できていないのではなく、そもそも機能を持たず、確率的にリン酸化されたに過ぎない部位も多いと提える化学量論的な報告もある。そこで、保存度との関連を議論するためにも、機能が特定されたリン酸化に限定して予測対象とする。

(4)データとして次のものを用いた。ヒトタンパク質のリン酸化データは、phosphoELM の version 9.0 (September 2010)より取得した。予測には、タンパク質当りのリン酸化部位数が 17 から 25 部位のタンパク質のみを用いた。ID 領域データは DICHOT より取得した。

４．研究成果

(1) アミノ酸配列情報を入力としサポートベクターマシン(SVM)を用いてリン酸化部位を予測した。その結果、ドメインでは IDR よりも高い精度を示した。その際、入力配列長 *Ws*、SVM のカーネルパラメータ、マージンパラメータを変えて、最も高精度のものを採用した。リン酸化対象となるアミノ酸残基の種別に、ドメインでは、セリンでは 77.8%(*Ws*=31)、トレオニンでは 74.7%(*Ws*=25)、チロシンでは 68.6%(*Ws*=31)であった。一方、ID 領域ではセリンでは 71.0%(*Ws*=7)、トレオニンでは 73.1%(*Ws*=9)、 チロシンでは 66.1%(*Ws*=7)となった。ドメインでは広範囲の配列情報が有効で高い精度を得るが、IDR では狭い範囲の配列情報しか有効でなく、十分な精度は得られなかった。これは、IDR の配列保存性の低さが原因と考えられる。

(2) IDR におけるアミノ酸残基ごとの進化的保存性を調べる方法として、他生物種のオーソログ配列と部位ごとに比較した。すなわち、解析対象のヒトタンパク質と他生物種のオーソログタンパク質の 2 本のアミノ酸配列をアラインメント後、同じ位置に同じアミノ酸残基が存在していれば、その部位が保存されていると定義する。ここでは、アラインメントには MAFFT を用いた。また、比較する生物種として、進化的距離と実験解析の豊富さか

ら、マウス、オポッサム、ニワトリ、ゼブラフィッシュの 4 つの脊椎動物を選択した。ヒトタンパク質データは UniProt(Release 2013_11)より、その他の生物種のタンパク質データは GTOP（Release 2010 October 6）より取得した。オーソログの判定には BLAST による双方向ベストヒットを用いた。それにより 4 生物種とのオーソログタンパク質が揃うヒトタンパク質のみに限定して、対象部位を以下の三つに分類した：

(A) 機能性リン酸化部位：UniProt にリン酸化部位であることが示されており、かつ、Description 欄にその明確な機能も判明している部位

(B) 非機能性リン酸化部位:リン酸化することは示されているものの、その明確な機能は記されていない部位

(C) 非リン酸化部位:リン酸化することが記されていない部位

その結果、IDR 上に機能性リン酸化部位を持つタンパク質は、124 タンパク質であった。その 124 タンパク質の IDR にある全てのセリンとトレオニンの部位数は、(A)223 部位、(B)527 部位、(C)6911 部位であった。(A)-(C)が、4 つの生物種それぞれにおいて保存されていた割合を図 1 に示す。全ての生物種で(A)、(B)、(C)の順に保存度が高いことが分かる。

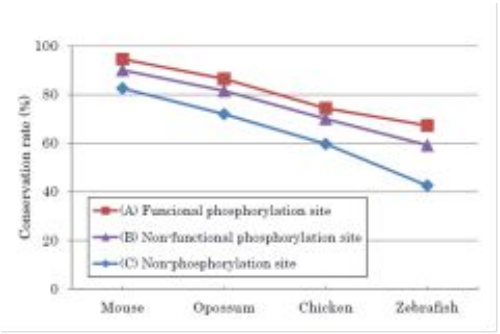

図 1　IDR での保存度

(3) SVM を用いて IDR における機能性リン酸化部位を予測した。SVM のポジティブデータとして、(A) 機能性リン酸化部位の全 223 部位を、ネガティブデータとして、(C) 非リン酸化部位の中からポジティブ部位と同数を等確率で選択した。

SVM への入力には、ヒト配列情報と MA 情報の 2 種類の頻度情報を用いて比較した。すなわち、前者は予測対象のヒトタンパク質のアミノ酸配列のみを入力する。それに対して後者は、ヒト、マウス、オポッサム、ニワトリ、ゼブラフィッシュの 5 本のオーソログタンパク質のアミノ酸配列を入力とする。5 つの生物種のアミノ酸配列をマルチプル アラインメントした後、各部位のアミノ酸残基、および、アラインメントで挿入されたギャップの出現割合を入力とする。ここではヒトタンパク質のギャップも、他の生物種と同様に扱い入力情報に含めた。

さらに、MA 情報と同様に進化的な情報を用いてリン酸化部位の予測を行っていると考えられる PPRED による予測と比較した。PPRED は、予測対象のタンパク質をクエリとして PSSM を取得し、それを入力情報として用いている。本研究では、予測対象のヒトタンパク質をクエリとし、UniProt(Release 2013_11)の全タンパク質データベースに対して PSI-BLAST を用いることで PSSM を取得した。このとき E-value は 0.001、イタレーション回数は 2 回と設定した。

上記それぞれの頻度情報に対して、BLOSUM62 の背景頻度を用いることでスコア化した情報による予測も行い比較した。

加えて、上記により学習した 2 つの SVM 用いたハイブリッド予測も行った。これは、2 つの SVM の予測が一致していた場合のみ、その予測を採用するものとし、一致していない場合は予測不可能として答えない。予測できた割合であるカバー率は下がるものの、高い精度での予測が期待できる。

SVM のパラメータを変えて得られた最大精度とそれを与えた配列長 *Ws* を表 1 に示す。背景頻度を考慮したスコア化により精度が向上していることが分かる。最も高い精度を与えたのは MA 情報、次に高い精度は PSSM 情報、最も低い精度となったのはヒト配列情報であった。

表 1　IDR における機能性リン酸化部位の予測精度

	Frequency	Score
Sequence	74.6% (Ws=11)	78.1% (Ws=9)
MA	77.7% (Ws=9)	79.2% (Ws=7)
PSSM	75.3% (Ws=9)	78.3% (Ws=7)
Hybrid	82.1%	-----

ヒト配列情報よりも PSSM や MA 情報が高い精度をとっていることから、進化的な保存度情報が予測に有効であることが分かる。また、PSSM 情報よりも MA 情報が高い精度での予測が行えている。PSSM は進化的な保存性の高いドメインでは有効に多くのアミノ酸配列を検索できる可能性があるものの、進化的な保存性の低い IDR では有効に働かない可能性がある。そのため、そうした IDR の PSSM 情報を用いた結果、有効な予測ができなかったと考えられる。

さらに、ヒト配列情報(頻度)と、MA 情報(スコア)によるハイブリッド予測の結果、それぞれ最大精度を与えた *Ws* の SVM の組み合わせで、82.1%の予測精度を達成し、そのカバー率は 81%であった。それぞれの SVM よりも高い精度を示しているため、これら 2 つは相補的な情報を含んでいると考えられる。

(4) ドメインでも同様に、SVM を用いた機能性リン酸化部位の予測を行い、IDR の結果

と比較した。

UniProt よりヒトタンパク質データを取得し、先と同じ 4 つの生物種でオーソログが揃う条件をつけた結果、ドメイン上に機能性リン酸化部位を持つタンパク質は、全 41 タンパク質であった。そのドメイン上の全対象部位を、以下の三つに分類した：

(A) 機能性離散化部位：50 部位
(B) 非機能性リン酸化部位：43 部位
(C) 非リン酸化部位：1968 部位

(A)-(C)が各生物種において保存されていた割合を図２に示す。ドメインであっても、(A)、(B)、(C)の順に保存度が高いことが分かる。

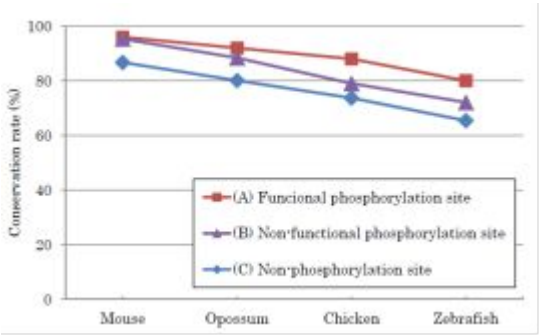

図２ ドメインでの保存度

(A)の機能性リン酸化部位である 50 部位をポジティブデータ、(C)の非リン酸化部位のうち同数をネガティブデータとして、SVM を用いてドメインでの機能性リン酸化部位の予測を行った。得られた最大精度と配列長 *Ws* を表２に示す。IDR では、MA 情報が最も高い精度となったのに対し、ドメインでは PSSM が最大精度をとり、ドメインでの予測では PSSM 情報が有効に働いていることが分かる。

表２ ドメインにおける機能性リン酸化部位の予測精度

	Frequency	Score
Sequence	73.8% (Ws=9)	74.7% (Ws=9)
MA	74.8% (Ws=9)	77.9% (Ws=15)
PSSM	76.8% (Ws=11)	79.0% (Ws=15)

(5) 対象部位周辺のアミノ酸出現頻度や出現の独立性を、(A) 機能性リン酸化部位と (C) 非リン酸化部位で比較した。IDR とドメインそれぞれで確認した。

ドメイン、IDR ともに、機能性リン酸化部位の P+1 が非常に多く、次いで R-3 も多く存在していた。いずれも既知だが、機能性リン酸化に限定するとより顕著であった。独立成分分析では、それら位置特異的に存在するアミノ酸残基は、互いに異なる独立成分となったことから、独立に出現していると思われる。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計４件）

① Xian-Hua Han, Yukako Tohsato, Koji Kyoda, Shuichi Onami, Ikuko Nishikawa, Yen-Wei Chen: Nuclear Detection in 4D Microscope Images of Developing Embryo Using Enhanced Probability Map of Top-ranked Intensity-ordered Descriptors, IPSJ Transactions on Computer Vision and Applications, Vol.8, No.8 (2016) 査読有
DOI: 10.1186/s41074-016-0010-3

〔学会発表〕（計１４件）

① 谷口元希、瀬尾昌孝、西川郁子：Deep 畳み込みニューラルネットワークによる実験データ分類の事例研究、第 60 回システム制御情報学会研究発表講演会、2016 年 5 月 27 日、京都テルサ（京都府・京都市）
② X.-H. Han, Y. Tohsato, K. Kyoda, S. Onami, I. Nishikawa and Y.-W. Chen: Nuclear Detection in 4D Microscope Images of Developing Embryo Using Enhanced Probability Map of Top-ranked Intensity-ordered Descriptor, The 3rd IAPR Asian Conference on Pattern Recognition, 2015 年 11 月 5 日, Kuala Lumpur (Malaysia)
③ I. Nishikawa, T. Ishino, Y. Tohsato, S. Fukuchi and K. Nishikawa: Prediction of Post-Translational Modification on Human Protein in Intrinsically Disordered Region by Support Vector Machine, 11th Neural Coding Workshop, 2014 年 10 月 8 日, Versailles (France)
④ 石野友喜、西川郁子、遠里由佳子、福地佐斗志、西川建：SVM によるタンパク質天然変性領域上の機能性リン酸化部位の予測、第 58 回システム制御情報学会研究発表講演会、2014 年 5 月 22 日、京都テルサ（京都府・京都市）

６．研究組織

(1)研究代表者
西川　郁子（NISHIKAWA, Ikuko）
立命館大学・情報理工学部・教授
研究者番号：９０２１２１１７

(2)研究協力者
西川　建（NISHIKAWA, Ken）
福地　佐斗志（FUKUCHI, Satoshi）
遠里　由佳子（TOHSATO, Yukako）