科学研究費助成事業　　研究成果報告書

平成　２６　年　　５　月　３１　日現在

機関番号：　３４３１５
研究種目：　基盤研究(C)
研究期間：　2011～2013
課題番号：　２３５００３７２

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修飾機構の解明

研究課題名（英文）Prediction of the post-translational modification sites of the protein by machine learning to study the modification mechanism

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）　　4,200,000円、（間接経費）　　1,260,000円

研究成果の概要（和文）：タンパク質の代表的な翻訳後修飾であるリン酸化を対象に、サポートベクターマシンを用いた機械学習により修飾部位を予測した。対象部位をドメインと天然変性(ID)領域に分けて取り扱った点が新規であり、ドメインではアミノ酸配列情報のみで十分予測可能だが、IDでは部位特異的な進化的保存度情報が有効であった。配列保存性が低いIDにおいては部位ごとの保存度は非一様であり、リン酸化部位、その中でも機能性が明確なリン酸化部位での保存度が高いことが分かり、翻ってそれらの予測に有効であった。

研究成果の概要（英文）：Phosphorylation is one of the most important post-translational modifications of the protein. Phosphorylation sites of the human protein are predicted using the machine learning by support vector machine (SVM). SVMs are constructed for prediction target sties in the functional domain and in the intrinsically disordered (ID) region, separately. As the result, human amino acid sequence information is enough for the effective prediction for the domain, while the site specific evolutionary conservation information turns out to be effective for ID. That is, conservation rate is not uniform in ID, where the sequence conservation is known to be relatively low in general. Site specific conservation is newly defined based on the ortholog proteins, and the conservation rate is high at the phosphorylation sites, especially at the functional phosphorylation sites, therefore which is effective for the prediction.

研究分野：　総合領域

科研費の分科・細目：　情報学・生体生命情報学

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

１．研究開始当初の背景

タンパク質への修飾とその機能の解明は、生命システムの理解に向けた主要課題のひとつである。申請者は、重要な翻訳後修飾でありながら、関与する酵素の多様性もあって実験的解析による機構の解明が困難でありコンセンサス配列の存否も不明だった *O* 型糖鎖修飾に対して、機械学習により部位を予測し成果を出していた。そこでは、修飾部位の密度という新規な視点で対象データを現象論的に分類することで予測精度を上げたが、それがタンパク質の天然変性(ID)領域と強く関連していることが分かった。そこで、他の翻訳後修飾やその中でも代表的なリン酸化に対して、最初から ID 領域に着目して機械学習を適用するという着想を得て研究を開始した。

リン酸化部位の予測に対する機械学習の適用は、様々な学習法や入力データを用いた先行研究が多いが、ID 領域との関連を陽に考慮したものはない。また ID 領域は、アミノ酸配列の進化的保存性が低いにも関わらず、多くの重要なリン酸化部位が存在しており、その理由も不明である。そこで、ID を陽に考慮した、リン酸化部位の予測を機械学習で実施することにより、ID 領域における予測精度の向上と、それを通した ID におけるリン酸化の特徴や機構の解明を目指した。

２．研究の目的

タンパク質の翻訳後修飾部位の機械学習による予測を通して、修飾機構の解明に結び付ける。申請者は先に、*O* 型糖鎖修飾に対して、サポートベクターマシン(SVM)による予測結果を元に、タンパク質の ID 領域との関連や、構造安定化・機能多様性における役割を調べた。同様の手法で、リン酸化に代表される他の翻訳後修飾反応や、全ゲノム配列解析が終了した哺乳類の全タンパク質に対して取り組む。すなわち、機械学習による予測を通じて修飾に関与する要因を絞り、修飾メカニズムや、修飾によりタンパク質が獲得する機能について、さらに、進化過程における獲得について仮説を立て、生物学的考察や実験により仮説検証を通して修飾機構を解明する。

（１）ドメインと ID におけるリン酸化を区別し、コンセンサス配列やモチーフなどの部位特異的なアミノ酸の存在で説明できるリン酸化と、進化的保存度に関わる要因を考慮すべきリン酸化など、諸特性の統計解析を通じて、必ずしも単一ではない修飾メカニズムを解明する。

（２）特に ID とリン酸化の関係に着目し、配列保存性が悪い ID におけるリン酸化部位の予測を通して、進化的背景との関係を明らかにする。

（３）精度のよい予測手法を確立し、未知修飾部位を予測する。また、実験による検証に結び付ける。

３．研究の方法

（１）ヒトタンパク質のリン酸化に焦点を絞り、対象部位の領域をドメインと ID に二分した上で予測を行う。リン酸化に絞る理由は、①最も研究が進んでいる修飾反応であり、②申請者が機械学習での成果を出した *O* 型糖鎖修飾と共通の性質が多く優位性があり、③タンパク質 ID 領域との関係が特に指摘されている点が挙げられる。

（２）　リン酸化データの取得

①　リン酸化部位を持つヒトタンパク質を PhosphoELM データベースから取得する。UniProt データベースのアノテーションとも照合する。

②　リン酸化部位の領域をドメインと ID に分ける。DICHOT, GTOP データベースから取得する。

③　リン酸化の機能が既報告か否かで対象部位を分ける。UniProt の Description から取得する。

（３）　部位ごとの進化的保存度の調査

①　ID におけるリン酸化部位に対して、進化的保存度を求める。ここで、ID は一般に配列保存性が低いため、保存度情報として機械学習でしばしば用いられてきた PSSM 情報は、アラインメントが不確実な場合には根拠が弱い。そこで、ID に対しても意味が明確な保存度として、対応部位が明確に同定できるオーソログタンパク質を用いて部位特異的保存度を定義する。

②　機能が明確なリン酸化部位を ID に持つヒトタンパク質に対して、マウス、ニワトリ、オポッサム、ゼブラフィッシュの４生物種のオーソログの有無を調査する。全４種にオーソログを有するタンパク質に限定し、５配列でマルチプルアラインメントを行い、リン酸化部位の保存度を求める。

（４）　配列と保存度情報を用いた予測。

①　機能が明確なリン酸化部位を正データ、リン酸化されない部位を負データとして SVM 学習器を構築する。機能性を考慮する理由は、機能性のないリン酸化の存在を化学量論的考察で説明する仮説を念頭におき、機能的なリン酸化部位こそを予測すべきとの方針による。

②　配列情報、保存度情報を用いて、教師あり学習により、予測器を構築する。予測に有効な入力情報をもとに、修飾に関与する要因、機構を解明する。

４．研究成果

（１）ドメインと ID に分けてリン酸化部位の予測器を構築することで、ドメインに対しては対象とするヒト配列情報のみで十分な精度(78%)を達成できることが分かった。約 80%の精度は、関連研究で有効性が主張される目安であり、リン酸化の活性度や持続時間、発生段階や組織などを区別せずに予測した場合の限界と考えている。セリン、スレオニンに対する予測器は、共通の学習を行うこと

も確認された。
（２）ID におけるリン酸化では、ヒト配列情報のみでは 70%程度の精度しか得られず、配列進化的保存性の低さと整合する。これに関連して計画外に展開した研究として、ID の統計解析から、タンパク質の進化的由来を解明した。ヒトのミトコンドリアに対して、ID の割合を解析することで、その由来を分類できることを示した。
（３）ID におけるセリンとスレオニンを、(A)機能が既報告のリン酸化部位、(B)機能が報告されていないリン酸化部位、(C)非リン酸化部位に分けて、部位特異的保存度を求めた。その結果、上記の順に保存度が高いことが分かった（図１）。即ち、ID 領域でも一様に進化的保存度が低いのではなく、修飾部位では高いことが分かる。さらに、機能を担わないリン酸化部位の存在が示唆される。

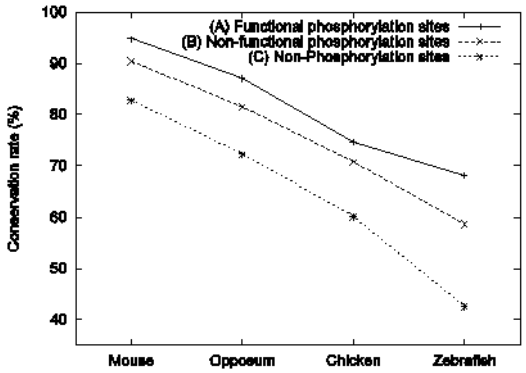

図１. ID 領域における Ser/Thr の保存度

（４）上記(A)と(C)の分類器を構築した。ヒト配列情報では 75%, ５生物種のマルチプルアラインメント配列情報では 77%, 両者を融合することで 80%の精度が得られた（図２）。リン酸化部位を機能性で分けること、進化的保存度情報を用いることの有効性が示された。

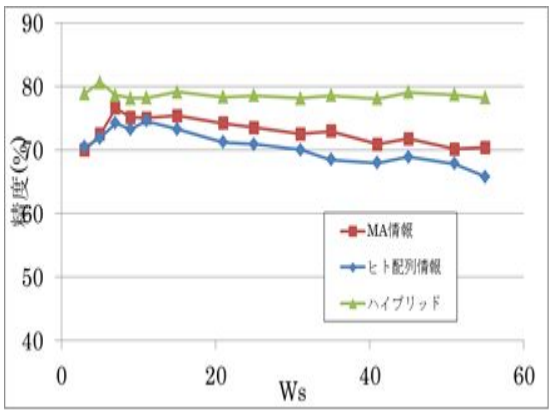

図２. 配列情報による SVM, MA 情報による SVM, および両者を組合せた予測の精度

（５）以上より、ID ゆえに部位特異的保存度の定義は従来困難であったが、単純で明確な定義のもとで定量的に求め、その有効性を、修飾部位の予測性能を通して示した。また、実験的に報告されているリン酸化部位に対して、機能性に着目する必要性を示した。さらに、機能性が未だ実験的には不明なリン酸化部位に対しても、ここで得られた予測器を適用することで、機能性の有無が予測可能となった。

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　８件）

① 著者名：R Kobayashi, S. Namiki, R. Kanzaki, K. Kitano, I. Nishikawa, P. Lansky、論文標題：Population coding is essential for rapid information processing in the moth antennal lobe、雑誌名：Brain Research、査読：有、巻：1536、発行年：2013、ページ：88-96、DOI：10.1016/j.brainres.2013.05.007
② 著者名：清水健史、榊原一紀、西川郁子、論文標題：トラック物流における配送日最適化問題と動的計画法に基づく近似最適化、雑誌名：システム制御情報学会論文誌、査読：有、巻：26、発行年：2013、ページ：365-373、DOI：10.5687/iscie.26.365
③ 著者名：H. Ikeno, T. Kazawa, S. Namiki, D. Miyamoto, Y. Sato, S. S. Haupt, I. Nishikawa, R. Kanzaki、論文標題：Development of a Scheme and Tools to Construct a Standard Moth Brain for Neural Network Simulations、雑誌名：Computational Intelligence and Neuroscience、査読：有、巻：2012、発行年：2012、ページ：ID 795291、DOI：10.1155/2012/795291
④ 著者名：M. Ito, Y. Tohsato, H. Sugisawa, S. Kohara, S. Fukuchi, I. Nishikawa, K. Nishikawa、論文標題：Intrinsically disordered proteins in human mitochondria、雑誌名：Genes to Cells、査読：有、巻：17、発行年：2012、ページ：817-825、DOI：10.1111/gtc.12000
⑤ 著者名：K. Sakakibara, Y. Tian, I. Nishikawa、論文標題：An Incremental Approach for Storage and Delivery Planning Problems、雑誌名：Decision Making in Manufacturing and Services、査読：有、巻：6、発行年：2012、ページ：5-23、DOI：10.7494/dmms/2012.6.1.5
⑥ 著者名：M. Zawidzki, K. Tateyama, I. Nishikawa、論文標題：The constraints satisfaction problem approach in the design of an architectural functional layout、雑誌名：Engineering Optimization、査読：有、巻：43、発行年：2011、ページ：943-966、DOI：10.1080/0305215X.2010.527005
⑦ 著者名：榊原一紀、田雅杰、西川郁子、論文標題：トラックターミナルを利用した配送・保管計画の整数計画モデルと数

理計画法による逐次的解法、雑誌名：システム制御情報学会論文誌、査読：有、巻：24、発行年：2011、ページ：88-96、DOI：10.5687/iscie.24.88

⑧ 著者名：池野英利、加沢知毅、並木重宏、シュテファン周一ハウプト、西川郁子、神崎亮平、論文標題：データベースを用いた脳・神経細胞データの管理と活用、雑誌名：比較生理生化学、査読：有、巻：28、発行年：2011、ページ：327-334、DOI：10.3330/hikakuseiriseika.28.326

〔学会発表〕（計　２８件）

① 発表者名：石野友喜、西川郁子、遠里由佳子、福地佐斗志、西川建、発表標題：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、学会名等：第５８回システム制御情報学会研究発表講演会、発表年月日：2014 年 5 月 22 日、発表場所：京都テルサ（京都府）

② 発表者名：T. Ishino, I. Nishikawa, S. Fukuchi, Y. Tohsato, K. Nishikawa、発表標題：Prediction of Protein Phosphorylation sites by Support Vector Machine、学会名等：The 6th International Conference on BioMedical Engineering and Informatics、発表年月日：2013 年 12 月 18 日、発表場所：Hangzhou（China）

③ 発表者名：I. Nishikawa, T. Ishino, Y. Tohsato, S. Onami, S. Fukuchi, K. Nishikawa、発表標題：Predicting Human Protein Phosphorylation Sites in Intrinsically Disordered Region by Support Vector Machine、学会名等：24th International Conference on Genome Informatics 2013、発表年月日：2013 年 12 月 17 日、発表場所：Biopolis（Singapore）

④ 発表者名：石野友喜、西川郁子、遠里由佳子、福地佐斗志、西川建、発表標題：SVM を用いたヒトタンパク質リン酸化部位の予測、学会名等：計測自動制御学会システム・情報部門学術講演会 2013、発表年月日：2013 年 11 月 19 日、発表場所：ピアザ淡海（滋賀県）

⑤ 発表者名：T. Ishino, I. Nishikawa, Y. Tohsato, S. Fukuchi, K. Nishikawa、発表標題：Prediction of Phosphorylation Sites by Support Vector Machine、学会名等：Innovation in Information and Communication Science and Technolgy 2013、発表年月日：2013 年 9 月 3 日、発表場所：Tomsk（Russia）

⑥ 発表者名：T. Kazawa, Y. Mori, D. Miyamoto, S. Namiki, H. Ikeno, S.S. Haupt, I. Nishikawa, R. Kanzaki、発表標題：Constructing a multi-compartment parallelized simulation of a premotor area of the silkmoth brain、学会名等：日本比較生理生化学会第３５回大会、発表年月日：2013 年 7 月 13 日、発表場所：イーグレひめじ（兵庫県）

⑦ 発表者名：R. Kobayashi, S. Namiki, R. Kanzaki, K. Kitano, I. Nishikawa, P. Lansky、発表標題：Decoding order identity from neural activity in the moth antennal lobe、学会名等：日本神経回路学会第２３回大会、発表年月日：2013 年 6 月 20 日、発表場所：国立京都国際会館（京都府）

⑧ 発表者名：石野友喜、西川郁子、榊原一紀、遠里由佳子、福地佐斗志、西川建、発表標題：ヒト蛋白質リン酸化部位の予測と保存性解析、学会名等：第１３回日本蛋白質科学会年会、発表年月日：2013 年 6 月 12 日、発表場所：鳥取県民会館（鳥取県）

⑨ 発表者名：西川郁子、発表標題：ニューロン生理データに基づく神経回路モデルの構築と検証、学会名等：電子情報通信学会複雑コミュニケーションネットワーク科学研究会、発表年月日：2013 年 6 月 4 日、発表場所：立命館大学（滋賀県）

⑩ 発表者名：石野友喜、榊原一紀、西川郁子、遠里由佳子、福地佐斗志、西川建、発表標題：サポートベクターマシンを用いた蛋白質のリン酸化部位予測、学会名等：第５７回システム制御情報学会研究発表講演会、発表年月日：2013 年 5 月 15 日、発表場所：兵庫県民会館（兵庫県）

⑪ 発表者名：石野友喜、榊原一紀、西川郁子、遠里由佳子、福地佐斗志、西川建、発表標題：SVM を用いたリン酸化部位予測と配列保存性の解析、学会名等：計測自動制御学会第４０回知能システムシンポジウム、発表年月日：2013 年 3 月 15 日、発表場所：京都工芸繊維大学（京都府）

⑫ 発表者名：I. Nishikawa, Y. Yamagishi, H. Ikeno, T. Kazawa, S. Namiki, R. Kanzaki、発表標題：Estimation of the Information Pathway in an Insect Brain based on the Physiological Data、学会名等：6th International Conference on New Trends in Information Science, Service Science and Data Mining、発表年月日：2012 年 10 月 24 日、発表場所：Taipei（Taiwan）

⑬ 発表者名：I. Nishikawa, Y. Yamagishi, H. Ikeno, T. Kazawa, S. Namiki, R. Kanzaki、発表標題：Estimation of the information pathway for a motor command generation in an insect brain based on the physiological data、学会名等：10th International Workshop Neural Coding 2012、発表年月日：2012 年 9 月 4 日、発表場所：Prague（Czech）

⑭ 発表者名：山岸嘉彦、小野島隆之、五十嵐吉輝、西川郁子、加沢知毅、並木重宏、池野英利、神崎亮平、発表標題：昆虫脳における生理応答データに基づく行動司令生成時の情報伝達経路の推定、学会名等：電子情報通信学会ニューロコンピューティング研究会、発表年月日：2012 年

7 月 31 日、発表場所：立命館大学（滋賀県）

⑮ 発表者名：百田直矢、加沢知毅、ステファン周一ハウプト、並木重宏、宮本大輔、神崎亮平、西川郁子、池野英利、発表標題：カイコガ標準脳データベース構築に向けた脳画像データの標準化と活用、学会名等：電子情報通信学会ニューロコンピューティング研究会、発表年月日：2012 年 3 月 16 日、発表場所：玉川大学（東京都）

⑯ 発表者名：西川郁子、小野島隆之、加沢知毅、並木重宏、池野英利、神崎亮平、発表標題：昆虫脳における運動司令生成のための神経回路の推定、学会名等：電子情報通信学会ニューロコンピューティング研究会、発表年月日：2011 年 12 月 20 日、発表場所：名古屋工業大学（愛知県）

⑰ 発表者名：I. Nishikawa, T. Igarashi, T. Kazawa, S. Namiki, H. Ikeno, R. Kanzaki、発表標題：Estimation of the Neural Circuit for the Command Generation in the Premotor Center of an Insect Brain、学会名等：6th International Conference on Computer Science and Convergence Information Technology、発表年月日：2011 年 11 月 30 日、発表場所：Jeju Island（Korea）

⑱ 発表者名：西川郁子、伊藤將弘、福地佐斗志、本間桂一、西川建、発表標題：機械学習によるタンパク質の *O* 型糖鎖修飾部位の予測とその機能の考察、学会名等：電子情報通信学会ニューロコンピューティング研究会、発表年月日：2011 年 7 月 26 日、発表場所：神戸大学（兵庫県）

⑲ 発表者名：山岸嘉彦、五十嵐吉輝、西川郁子、加沢知毅、並木重宏、池野英利、神崎亮平、発表標題：カイコガ前運動中枢ニューロンの生理応答の分類、学会名等：第５５回システム制御情報学会研究発表講演会、発表年月日：2011 年 5 月 18 日、発表場所：大阪大学（大阪府）

〔図書〕（計　０件）

〔産業財産権〕
○出願状況（計　０件）

○取得状況（計０件）

〔その他〕
ホームページ等
http://www.sys.ci.ritsumei.ac.jp/

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)研究分担者
なし

(3)連携研究者
なし