2005－BIO－3（1）
2005／12／22

機械学習によるタンパク質 N-ミリストイル化規則の予測

岡田 諒 † 杉井 学 †† 松野 浩嗣 ††† 宮野 悟 ††††
† 山口大学大学院 理工学研究科
†† 山口大学 メディア基盤センター
††† 山口大学 理学部
†††† 東京大学 医科学研究所 ヒトゲノム解析センター

タンパク質 N-ミリストイル化は、真核生物及び、ウイルス由来のタンパク質の N 末端に炭素数 14 の飽和脂肪酸であるミリスチン酸が共有結合するタンパク質の脂質修飾である。本研究では機械学習システム BONSAI を用いて、タンパク質 N-ミリストイル化に対する配列要求を調べた。BONSAI は、インデキシングと決定木の形で規則を発見する。実験の結果、BONSAI はミリストイル化に影響すると考えられているアミノ酸をインデキシングした。さらにミリストイル化を誘導する配列に関して、従来より重要とされてきた位置に加え、新たに、関係ないとされていた位置についてのアミノ酸要求の可能性も示唆した。

Machine Learning Prediction of Amino Acid Sequence Characterization in Protein N-Myristoylation

Ryo Okada† Manabu Sugii†† Hiroshi Matsuno††† Satoru Miyano††††
† Graduate School of Science and Engineering, Yamaguchi University
†† Media and Information Technology Center, Yamaguchi University
††† Faculty of Science, Yamaguchi University
†††† Human Genome Center, Institute of Medical Science, University of Tokyo

Protein N-myristoylation is the lipid modification in which the saturated fatty acid of 14 carbon number binds covalently to N-terminal of virus-based and eukaryotic protein. In this study, we suggest an approach to predict the pattern of N-myristylation signal using the machine learning system BONSAI. BONSAI finds rules in combination of a alphabet indexings and decision trees. The experiment showed that BONSAI classified the amino acid according to effect for N-myristoylation and found the rules in the alphabet indexing. In addition, BONSAI suggested new requirements for the position of amino acid in the N-myristoylation signal.

1　はじめに

タンパク質 N-ミリストイル化は、真核生物及び、ウイルス由来のタンパク質の N 末端に炭素数 14 の飽和脂肪酸であるミリスチン酸が共有結合するタンパク質の脂質修飾である。ヒトゲノムがコードする全タンパク質の約０．５％にこの修飾が生じているものと推定されており [1]、ミリストイル化によって、細胞膜の情報伝達など、多様な生理機能を実現することが分かっている [2][3]。ミリストイル化を生じるタンパク質の N 末端にはミリストイル化を指令する、N ミリストイルシグナルと呼ばれる配列が存在する。

この配列は N 末端から 6 から 9 アミノ酸程度と考えられている。これまでこのミリストイル化を正確に予測することを目標に、その配列要求が調べられてきた [4][5]。ただその手法は、研究者がそのパターン発見を生物学実験を基にして、培ってきた知識から配列を見ることによって、そのパターンを予測するというものであった。しかしその配列情報は膨大であり、さらにそこには一つの単一的な規則ではなく、特異的な例外も非常に多く含まれる。そのため今後は計算機を用いた、情報科学的手法による大量のデータから規則を予測するということが、ミリストイル化規則の予測においても重要となってくる。

機械学習システム BONSAI は、そういった複数のアミノ酸配列などの一次構造データから、知識を獲得することができるシステムである [6]。BONSAI は正の例と負の例からそれらを分かつ規則を決定木の形で発見する。また、それと同時に、その決定木を構成するパターンを作る上で便利なように、複数の文字を一つの別の文字に置き換える、インデキシングという作業も行なう。

今回我々は BONSAI を使い、ミリストイル化に特有なアミノ酸配列を発見することを目的として、計算機実験を行った。

2. では、タンパク質 N-ミリストイル化について、その配列要求に重点を置き、その特徴について述べる。3. では、ミリストイル化規則を発見するのに用いた、機械学習システム BONSAI についてその動作と、特徴について述べる。4. では、BONSAI を用いたミリストイル化規則の発見の実験について、本実験での手法について述べる。5. では、BONSAI が発見した規則について述べる。本実験では 2 つの興味深い規則があらわれたが、ここではその解釈と妥当性について詳しく見ていく。

2 タンパク質 N-ミリストイル化

タンパク質 N-ミリストイル化は、真核生物および、ウイルス由来のタンパク質の N 末端に炭素数 14 の飽和脂肪酸であるミリスチン酸が共有結合するタンパク質の脂質修飾である。ヒトゲノムがコードする全タンパク質の約 0.5 %にこの修飾が生じているものと推定されている。

ミリストイル化は、リボソーム上におけるタンパク質の翻訳途中にメチオニンアミノペプチダーゼにより開始メチオニンが切断除去され、露出した N 末端グリシン残基の α-アミノ基に N-ミリストイル転移酵素（NMT）がミリストイル CoA のミリストイル基を転移することにより生じる（図 1）。

その結果生じたミリストイル化タンパク質の多くは、細胞情報伝達に直接関与する生理活性タンパク質であり、細胞膜やオルガネラ膜との結合を介して固有の機能を発現する。ミリストイル化を介した膜への結合はきわめて多様な制御を受け、細胞の情報伝達やウイルスの増殖過程においてタンパク質の機能調節機構に重要な役割を演じていることが明らかになっている [2][3]。例えば、HIV-1 の Gag タンパク質は、N 末端ミリストイル基を利用して、原形質膜へと移行し、原形質膜上でウイルス粒子形成や出芽に関与する。また、アポトーシス誘導因子 Bid は、細胞質中でのプロテアーゼによる切断に伴い、切断後に新しく生じた N 末端にもミリストイル化が生じることが明らかになっている [7]。

ミリストイル化を生じるタンパク質の N 末端には、ミリストイル化を指令する特異的な配列が存在する。通常この配列は 6 から 9 アミノ酸程度と言われており、長くても 17 アミノ酸程度だと考えられている [1]。しかし N 末端から距離が離れるほど、その影響は弱くなる。ミリストイル化が起きるタンパク質の N 末端配列の一例を、表 1 に示す。

これまでの研究から、この配列には、N 末端の開始 Met に続いて Gly 残基が必須であり、その次に位置する 3 位と 6 位のアミノ酸が修飾反応に大きく影響することが明らかになっている。さらに最近の研究 [4] 及び [5] から、図 2 に示すように 6 位が Ser の場合は、11 種の

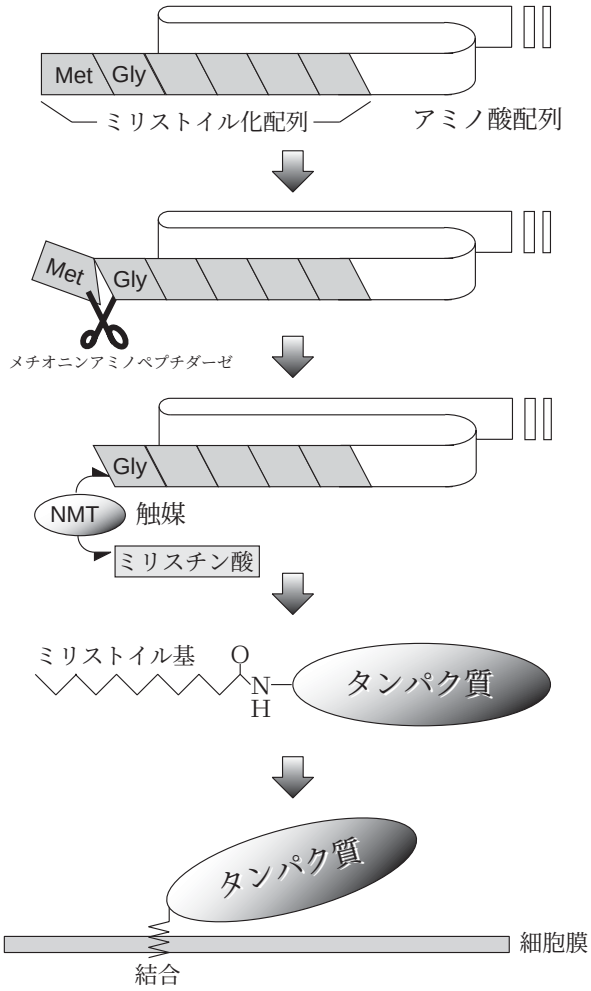

図 1: タンパク質 N-ミリストイル化

アミノ酸が 3 位に存在するときに効率的なミリストイル化が生じることが明らかになっている。また、この 3 位が Ser の場合に 6 位で許容される 11 種類のアミノ酸には規則性があり、そのほとんどが回転半径 1.80 Å以下のアミノ酸である。実際にこれより回転半径の大きいもの全てが 6 位で許容されない。例外として Pro、Asp、Glu があるが、例えば Pro などはその分子的な性質上、ミリストイル化に重要なほとんどの部位で許容されないことも分かっている。

一方、Ala が 6 位に存在する場合は、5 種類のアミノ酸が、Thr が Phe が 6 位に存在する場合、わずか 2〜3 種類のアミノ酸が 3 位に存在するときにのみ効率のよいミリストイル化が生じることが明らかになっている。また、6 位以外に 7 位のアミノ酸が 3 位のアミノ酸の要求性に影響を与える場合も存在する。通常 6 位が Ser の場合 3 位で Lys は許容されないが、例外的に 7 位に Lys が存在すると 3 位のアミノ酸要求が変化し、3 位で Lys

図 2: タンパク質 N-ミリストイル化規則

表 1: ミリストイル化配列の例

タンパク質	アミノ酸配列
GAG SIVM1	MGARNSVLSGKKADE
KCRF STRPU	MGCAASSQQTTATGG
Q26368	MGCNTSQELKTKDGA
GBAZ HUMAN	MGCRQSSEEKEAARR
GAG MPMV	MGQELSQHERYVEQL
Q67940	MGQNLSTSNPLGFFP
COA2 POVM3	MGAALTILVDLIEGL
COA2 SV40	MGAALTLLGDLIATV
RASH RRASV	MGQSLTTPLSLTLDH
BASP BOVIN	MGGKLSKKKKGYNVN

が許容されるようになることも見出されている [5]。

3 機械学習システム BONSAI

機械学習システム BONSAI は 1 次元の記号列データからの知識獲得のための機械発見システムである [6]。図 3 にこのシステムの構成図を示す。このシステムは、図 4 のように正の例と負の例からなる記号列の集合が与えられると、それらを分類する仮説として、アルファベットのインデキシングと決定木を提示する。この決定木でのパターン上で用いられている記号はインデキシングにより変換されたものである。

インデキシングとは、記号列データの要素をグルーピングし、種類を減らす作業のことである。例えば 7 種類の疎水性のアミノ酸が 3 つ連なっている、というパターンを検索する場合を考える。このパターンを記述する場合、通常であれば 7 の 3 乗通りのパターンを記述し、検査しなければならない。だがこれら 7 つの疎水性のアミノ酸を 1、そうでないアミノ酸を 0 という風にインデキシングを行なった場合、「１１１」という 1 パターンのみを検索することによってそれを判断することが可能となる。そしてこうしてインデキシングを行なったパターンに対して、BONSAI は決定木を作成する (図 5)。

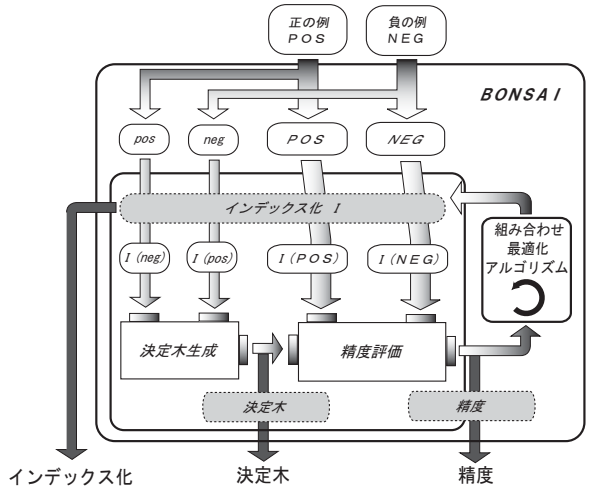

図 3: BONSAI の構成

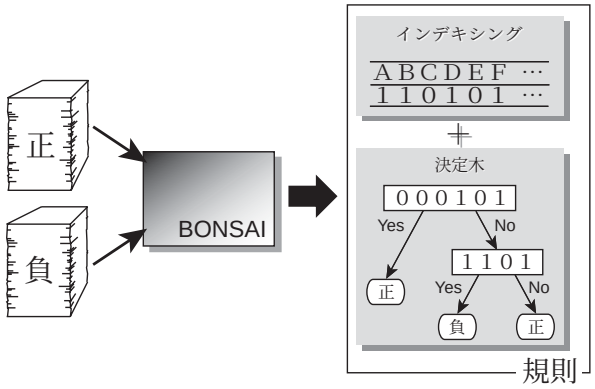

図 4: BONSAI の動作

インデキシングによって、計算の高速化と規則の表現を容易にすることが可能である。

このようなインデキシングという作業を、BONSAI は決定木作成と同時に自動的に行なう。それによって、規則の表現を柔軟にするだけでなく、利用者が思いつかなかった新たな知識が発見されることが実証されている。その例として実際に BONSAI を使った実験の一つに、タンパク質の膜貫通領域を予測する実験 [6] が行なわれているが、この実験で BONSAI は、親水度が低いアミノ酸を 1、親水度が高いアミノ酸を 0 と置き換え、投入した例の 90 %以上を分ける決定木を提示した。

4 アミノ酸部位を特定した配列パターンの発見

本実験では、生化学実験によりミリストイル化が起きることが明らかになっている配列を正の例とし、そうでない配列を負の例として BONSAI に投入し、ミリスト

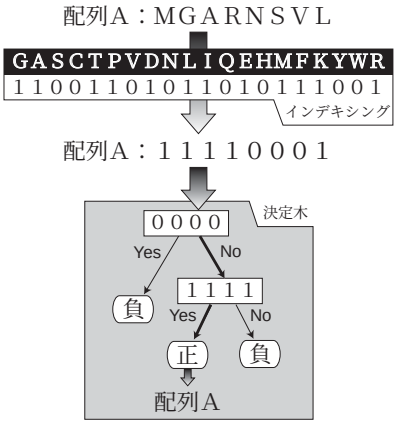

図 5: インデキシング

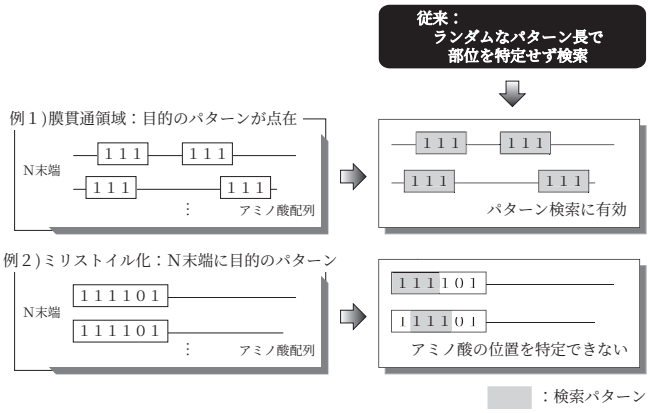

図 6: 従来の BONSAI のパターン検索

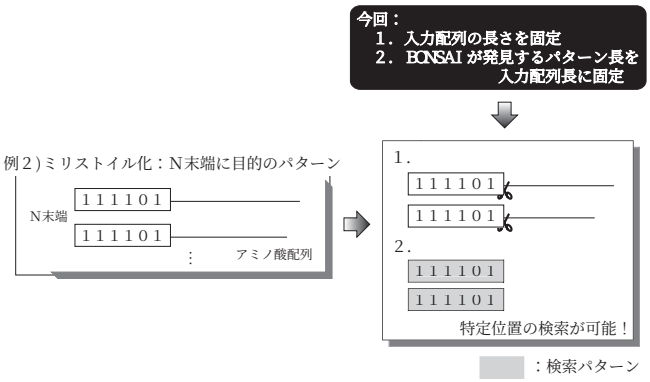

図 7: 今回の BONSAI のパターン検索

イル化の規則を発見した。なお、本実験では、ミリストイル化が起きるとされている配列 78 本と、文献 [2] 中で効率的にミリストイル化が起きることが確認されている配列を正の例とし、負の例には NCBI[8] データベースより得たヒトの全タンパク質配列からランダムに選択した。今回、負の例をヒトのタンパク質中からランダムに選択したが、これはヒトゲノム中でミリストイル化される配列は 0.5 %程度に過ぎないと推定されていることからである。

今回の実験では、入力するアミノ酸配列の長さや、20 種類のアミノ酸をインデキシングする文字数の適切な値を決定するという目的から、これらの値を様々に変えて BONSAI に投入し、実験を行った。この際、N 末端から 1 位についてはその全てが Met であるため、除外してある。

また今回は BONSAI を改変して、ノードで判定されるパターン長を全て一定にして実験を行なった。本来 BONSAI は、決定木の枝にあたる検索するパターン長は一定ではなく、図 6 のように特定のパターンが、目的の配列内に存在するかということだけを判断する決定木を作成する。そのため BONSAI が発見したパターンは、配列中のどの部位、第何位にそのパターンがあるのかということは考慮されない。しかしこの方式では、先の膜貫通領域の実験 [6] のように、目的とされるパターンが入力された配列に点在するという場合ならよいが、今回のミリストイル化のように、N 末端から何位のアミノ酸が重要であるというような、特定の位置に着目した規則を発見することは難しい。例えば、今回の N ミリストイル化は N 末端の 1 、 2 位に Met、Gly という配列が存在する必要性があるが、BONSAI がそのような配列が重要であるという規則を見つけたとしても、調べられるのは配列内に Met、Gly という配列が存在するかということだけであり、それがそれは N 末端の 3、4 位であるかもしれず、1、2 位であるとは限らない。

そこで今回、図 7 のように、まず投入するアミノ酸配列の長さを一定とし、さらに BONSAI が発見するパターン長を、その投入するアミノ酸配列と全て同じ長さにするように BONSAI を改変して実験を行なった。こうすることによって、投入するアミノ酸配列の全ての部位を特定できる規則が発見される。つまり、例えば、投入するアミノ酸配列の長さが 20 で、BONSAI によって発見されるパターン長が同じ 20 であるなら、投入したアミノ酸の 1 位は発見されたパターンの 1 位、2 位は 2 位というように対応しており、アミノ酸の第何位に注目しているのかということを明確に表現した規則を得ることができる。

5 N-ミリストイル化規則の予測

BONSAI によって求められた、いくつかの結果の中から、図 8 及び図 9 に示すような 2 つの興味深い結果が得られた。一つは既存の規則の確認となるような結果が表現されており、もう一方では既存の規則と未知の規

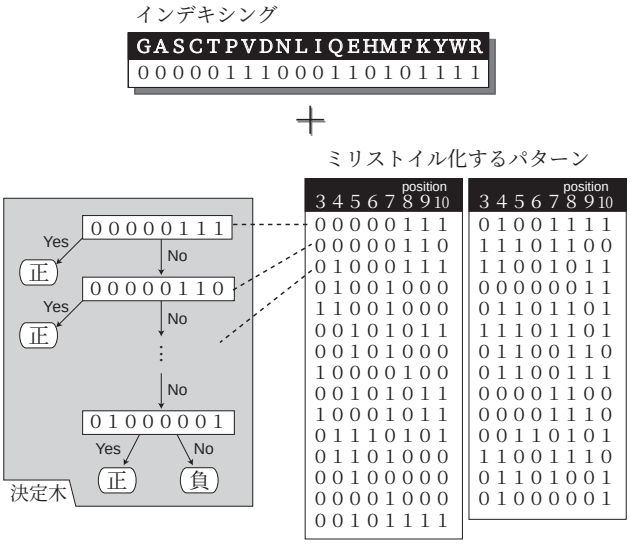

図 8: 規則 1 の決定木とインデキシング

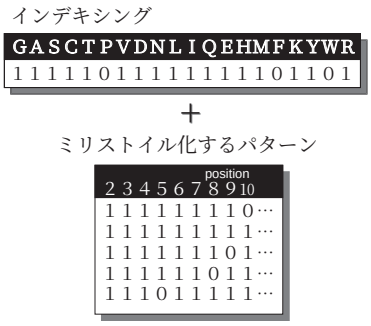

図 9: 規則 2 の決定木とインデキシング

則が表現されていた。ここでは前者を規則 1、後者を規則 2 とし、それぞれの特徴と考察を以下に述べていく。

5.1 規則 1：既存の規則の確認

正の例として、従来よりミリストイル化されることが分かっている配列に、論文 [2] でミリストイル化されるとした配列を加えた 91 の配列を用いた。負の例には、NCBI データベースよりランダムに選択した 800 のヒトの配列を用いた。また、N 末端から 1 位を除いて 9 残基のアミノ酸についての規則を調べた。

図 8 パターンを見てみると、特定の位置に特異的な偏りが生じていた。つまり 39 個のパターンのうち、3 位は 30 個で 0、6 位では 32 個で 0 の場合に、ミリストイル化が起きるという規則があらわれていた。まず 6 位については、今回入力した正の例のほとんどが Ser であることから説明できる。そこで 3 位に注目した場合、0 デインデキシングされているアミノ酸は、6 位が Ser の

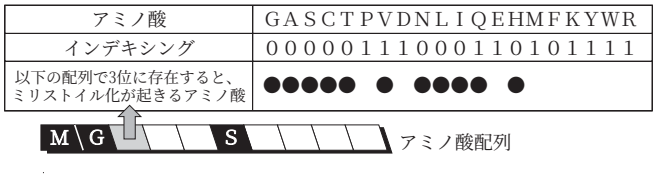

図 10: 規則 1 のインデキシング

場合に 3 位で許容されるアミノ酸であった。前述したように、6 位が Ser の場合、3 位に 11 種のアミノ酸が存在したときに効率的なミリストイル化が生じる。この結果では、図 10 に示すように、その 3 位で要求される 11 種のアミノ酸とそうでないアミノ酸のほとんどを、BONSAI がインデキシングで分けていた。つまりミリストイル化が起きる 11 種のうちの 9 種を 0、ミリストイル化が起きない 9 種のうちの 8 種を 1 としていた。

ところで、決定木を見てみると、3 位が 1 であるものもいくつか存在した。しかし、図 8 に示すように、これら 3 位が 1 である決定木は必ず 7 位が 1 であった。これは、先に述べていた本来 3 位に Lys は許されないが、7 位が Lys である場合に、特異的に 3 位で Lys が許されるという性質を表現していると考えられる。

5.2 規則 2：既存の規則と未知の規則の発見

これは従来よりミリストイル化されることが分かっている 78 の配列を正の例、NCBI データベースよりランダムに取ってきた 100 のヒトの配列を負の例として、N 末端から 1 位を除いた 19 残基のアミノ酸についての規則について調べた。ただし、11 位以降についてはミリストイル化に強く影響を与えないと考えられているため [1]、今回は省略している。

結果として、「ある配列がミリストイル化するならば、その配列中に Pro、Phe、Try のアミノ酸が 5、8、9、10 位に 1 つだけ存在するか、もしくはそれらのアミノ酸が配列中に全く存在しない」という規則が得られた。この規則は対偶をとることによって「あるタンパク質について Pro、Phe、Try が 5、8、9、10 位に 2 つ以上存在する場合、もしくはこれらアミノ酸が 2、3、4、6、7 位にあった場合、そのタンパク質は N-ミリストイル化が起きない」という規則に書き直せる（図 11）。

まずこの書き直した規則の、「Pro、Phe、Try が 5、8、9、10 位に 2 つ以上存在するならば、ミリストイル化しない」について見ていく。現在その 8、9、10 位についてはミリストイル化に対するアミノ酸が定義されておらず、また 5 位などは、どのようなアミノ酸が入ってもミリストイル化には影響を与えないとされている [3]。しかし本実験では、この 5 位を含め、2ヶ所以上に Pro などのアミノ酸が入ることによって、ミリストイル化が阻

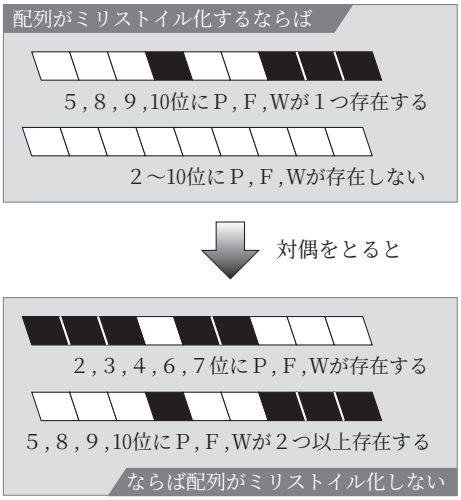

図 11: 規則 2 の解釈

害されるという結果を提示している。これは、ここで注目されている Pro などはタンパク質の 3 次構造に大きな影響を与えることが知られており、それが 2 箇所以上の場所に入ることによって、タンパク質の構造が大きく変わってしまうということを表現しているものと考えられる。

次に後半の「Pro、Phe、Try が 2、 3、4、6、7 位にあった場合に、ミリストイル化が起きない」という箇所であるが、2、3、6、7 位などの重要な場所については、すでに Pro などのアミノ酸が許容されないことが分かっており、これらの結果は、従来の規則と一致している。さらに 4 位については、これまでどのようなアミノ酸が存在してもミリストイル化が起きるとされてきたが、ここでもそれら 3 つのアミノ酸は許容されないという、新たな可能性を示唆していた。

6 おわりに

本研究では機械学習システム BONSAI を用いて、タンパク質 N-ミリストイル化に対する配列要求を調べた。BONSAI は、ミリストイル化に影響すると考えられているアミノ酸でインデキシングを行った。さらにミリストイル化を誘導する配列に関して、従来より重要とされてきた位置に加え、新たに関係ないとされていた位置についてのアミノ酸要求の可能性も示唆した。

今後は、従来のミリストイル化が起きるとされている配列とは異なる結果に注目し、生化学的手法を用いた実験とともに検証していきたい。また、今回は発見された個々の規則について個別に着目するということはせず、それら規則の大まかな特徴について見ている。そのため、ミリストイル化が起きる配列中に許容される、部位特異的なアミノ酸の規則について検討を行なっていないが、これについても今後検討していきたいと考えている。さらに、今回は BONSAI のパターン長を一定にするという手法で実験を行なったが、このことにより BONSAI がノイズとなる部分の規則も発見しようとして、規則の精度を下げているという場合も見受けられた。そのため、今後はこれに対応できるような BONSAI の利用法の検討も進めていきたい。

本実験で、ミリストイル化のような配列要求が明らかになっていない配列に対して、機械学習システム BONSAI を用いた配列予測が有効であることが明らかになった。さらに同様の配列検索によって、他のシグナル配列の予測などに対しても機械学習システム BONSAI は有効であると考えられる。

謝辞

本実験を行なうにあたり、山口大学農学部・内海俊彦教授には、多大なご教示やご示唆をいただいた。ここに記してお礼を申し上げる。

参考文献

[1] Maurer-Stroh S., Eisenhaber B., Eisenhaber F., “N-terminal N-myristoylation of proteins prediction of substrate proteins from amino acid sequence”, *J. Mol. Biol.*, 317(4):541-557, 2002.

[2] Resh, M. D., “Fatty acylation of proteins: new insights into membrane targeting of myristoylated and palmitoylated proteins”, *Biochim. Biophys. Acta*, 1451, 1-16

[3] Farazi, T. A., Waksman, G. and Gordon. J. I., “The biology and enzymology of protein N-myristoylation”, *J. Biol. Chem.*, 276(43):39501–39504, 2001.

[4] Utsumi, T., Sato, M., Nakano, K., Takenuma, D., Iwata, H., “Amino Acid Residue Penultimate to Amino-terminal Gly Residue Strongly Affects Two Cotranslational Protein Modifications, N-Myristoylation and N-Acetulation”, *J. of Biol. Chem.*, 276(13):10505-10513, 2001.

[5] Utsumi, T., Nakano, K., Funakoshi, T., Kayano, Y., Nakao, S., Sakurai, N., Iwata, H., and Ishisaka, R., “Vertical-scanning mutagenesis of amino acid in a model N-myristoylation motif reveals the major amino-terminal sequence requirements for protein N-myristoylataion”, *Eur. J. Mol. Biochem.*, 271(4):863–74, 2004.

[6] Shimozono, S., Shinohara, A., Shinohara, T., Miyano, S., Kuhara, S., Arikawa, S., “Knowledge Acquisition from Amino Acid Sequences by Machine Learning System BONSAI”, *Trans. Inform. Process. Soc. Japan*, 35(10):2009-2018, 1994.

[7] Zha, J., Weiler, S., Oh, KJ., Wei, MC., Korsmeyer, SJ., “Posttranslational N-myristoylation of BID as a molecular switch for targeting mitochondria and apoptosis”, *Science*, 290(5497):1761-1765, 2000.

[8] NCBI,
$ftp://ftp.ncbi.nih.gov/$