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　　５　日現在

機関番号：　３２６９０
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2016
課題番号：　２６３３０３３３

研究課題名（和文）糖鎖機能解明のための機械学習モデル開発及び糖鎖プロファイルデータベースの構築

研究課題名（英文）Development of a machine learning model and glycan profile database for understanding glycan function

研究代表者

木下　聖子（Aoki-Kinoshita, Kiyoko）

創価大学・理工学部・教授

研究者番号：５０４４０２３５

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：糖鎖機能解明のための機械学習モデル開発にあたり、事前に学習するモデルのトポロジーを決定する必要があり、そのため、入力された糖鎖構造のマルチプルアラインメントを行うアルゴリズムを開発することになった。MCAWツールを開発し、これ自体を用いて糖鎖の機能解明が可能であることが明らかになり、MCAWを用いた解析を行うことにした。解析には糖鎖アレイなどの技術から得られた糖鎖と糖鎖を認識するタンパク質間の結合親和性情報を用い、MCAWで解析した結果、文献と一致する結果を得ることができた。今後は解析した結果をデータベース化し、糖鎖のプロファイルを検索可能にする予定である。

研究成果の概要（英文）：In order to develop a machine learning model for understanding glycan function, it was necessary to first determine the topology of the pattern to learn from glycan data. Thus, it became necessary to develop a multiple glycan alignment algorithm, which we called MCAW. From this we realized that it was possible to use MCAW directly to analyze glycan function from technologies such as glycan arrays. Therefore, we proceeded to analyze such experimental data using MCAW. We used the high-affinity glycans of glycan-binding proteins as input to MCAW. As a result, we found that MCAW was able to produce glycan recognition patterns that have been confirmed in the literature, in addition to other patterns that were not necessarily found readily. Thus we show that MCAW was able to produce results based on the experimental data. We are working on developing a glycan profile database based on these analytical results.

研究分野：　糖鎖インフォマティクス

キーワード：　糖鎖　アラインメント　ウェブツール　糖鎖認識タンパク質　糖鎖インフォマティクス

１．研究開始当初の背景

米国の Consortium for Functional Glycomics (CFG)が糖鎖アレイの解析結果をウェブ上で公開した結果、糖鎖の認識されるパターン解析研究が促進した。これらのデータを用いて、米国を中心に Smith et al. 2010 や Cholleti et al. 2012 等により統計的なパターン解析が行われた。彼らは、結合する糖鎖と結合しない糖鎖の比較によって糖鎖の認識される部分構造の抽出を試みた。しかし、糖鎖構造の情報には様々なノイズが存在する。アレイ上の糖鎖にはコア構造が無かったり、非還元末端の部分構造しか無かったりし、結合情報も不完全である場合も多い。また、糖鎖構造の曖昧性が解析を困難にしている。例えば、ガラクトースと N アセチルガラクトサミンが同様の単糖として扱われたり、糖結合の種類の間にも類似性が見られたりすることも知られている（Aoki et al., 2005）。

そこで、我々は機械学習を用いて確率モデル Profile Probabilistic Sibling-dependent Tree Markov Model (ProfilePSTMM)（Aoki-Kinoshita et al., 2006）を開発した。このモデルは、糖鎖のパターンをプロファイルとして出力ができるため、曖昧な糖鎖のパターンを抽出することができる。また、ノイズを含むデータにも対応でき、糖鎖特有のパターンも学習できる。糖鎖を「木」としてみた場合、「兄弟」関係を考慮し、糖鎖認識に特徴的なパターンを抽出する。

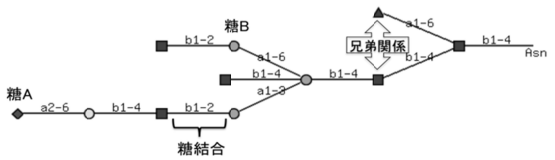

図１：糖鎖の木構造は糖結合の種類によって順序付きの兄弟関係を含む。

しかし、このモデルが学習するプロファイルの形（ステートモデル）を学習前に決める必要がある。つまり入力された糖鎖構造を基に自動的にステートモデルを決めるアルゴリズムが必要になった。そこで、我々は糖鎖構造のマルチプルアラインメントのアルゴリズム(MCAW)及びツールを開発した（Hosoda et al., 2012）。MCAW はアラインメントのスコアを計算するために様々なパラメータを要する。生物学的に妥当なアラインメントを得るためのパラメータを設定し、CFG のアレイデータの解析を行った結果、文献の結果と一致するアラインメントを得ることができた。

２．研究の目的

当初の目的は MCAW を用いて ProfilePSTMM のステートモデルを決定し、新たに高効率な確率モデルを開発することであった。またこの新たなモデルで CFG の糖鎖アレイデータベースを解析し、糖鎖の認識されるプロファイルデータベースを構築することであった。しかしながら、途中で MCAW においての難点が生じて、MCAW のアラインメント精度を確認する必要があり、MCAW を用いた解析に多くの研究時間を要した。そこで、本研究の目的を、確率モデルを開発する前段階の糖鎖のマルチプルアラインメントの解析を行うことにした。

３．研究の方法

まず、MCAW の有用性を確認するために、意図的に共通する糖鎖モチーフを組み込んだ糖鎖構造のデータセットを作成し、モチーフがアラインメントされることを確認する。この時、選んだ糖鎖モチーフは sialyl Lewis X 構造 (Neu5Ac α 2-3Gal β 1-4(Fuc α 1-3)GlcNAc) で、この部分構造を持つ 10 個の糖鎖構造を入力し、MCAW のデフォルトのパラメータで実行した。

次に、CFG および Lectin frontier Database (LfDB)のデータベースに糖鎖と糖鎖認識分子との間の結合親和性を測った実験結果が格納されている。これらのデータベース化ら高親和性糖鎖構造を収集し、それぞれの分子に対する認識パターンを確認する。具体的な手順として、まず MCAW で解析する CFG 及び LfDB のデータセットを作成する。CFG では糖鎖アレイデータを用いて実験されており、糖鎖と糖鎖結合タンパク質との親和性が蛍光強度 RFU 値で表されている。信頼の高い結合強度を示す糖鎖構造をデータセットに含ませるために、一番親和性が高い RFU 値を 100 とし、それをもとに他の RFU 値を計算する。計算した値が 75 以上かつ RFU 値の変動係数が 20 未満になる糖鎖構造をデータセットに採用する。LfDB ではフロンタルアフィニティクロマトグラフィーを用いた実験のデータが蓄積されており、レクチンと相互作用を示す糖鎖構造データをデータセットとして使用する。そして解析した結果をまとめて、データベース化する。

なお、MCAW の弱点は分岐点にギャップを挿入してアラインメントを作成することができない点であった。そのため、MCAW のアルゴリズムとは異なる、編集距離を用いた糖鎖のマルチプルアラインメントアルゴリズムを実装したツールも新たに開発することとした。

４．研究成果

構造中に類似するモチーフを意図的に組込んだデータセットを MCAW で解析した結果、モチーフが高いパーセンテージでアラインメントされていることが確認できた。様々な糖鎖構造中の共通部位のアラインメント結果を得られることから、MCAW の有用性が確認できた（図２）。

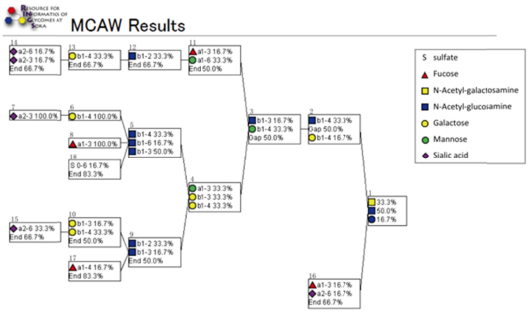

図２：sialyl-Lewis X 構造を含む MCAW の結果において、位置 5～8 に 100%で一致した。

さらに、実際に糖鎖–糖鎖認識タンパク質の相互作用データを使用し、解析を行った。CFG の糖鎖アレイデータと LfDB のフロンタルアフィニティクロマトグラフィーの実験結果を用いて、それぞれのデータセットに含む糖鎖構造に対して結合親和性の強度に相当するよう、入力した糖鎖構造の数を合わせた。高い親和性のある糖鎖構造の数を低いものと比例して数を増やした。その結果、90 以上のアラインメント結果が得られた。

得られた解析の内、濃度が異なった、あるいは実験が違う同じ糖結合タンパク質から得られた実験データのアラインメント結果に注目すると、高い割合でアラインメントされた糖鎖構造プロファイルは同じ形をとることも確認できた（Hosoda et al., 2017）。これは、認識結合部位を解析する上で必要な着目点の１つに相当すると考えられる。実験データが蓄積されている糖鎖結合タンパク質の濃度は、様々であり、実験者によって異なる。従って、同じタンパク質に対して類似するアラインメントが得られることがわかった。

また、我々が行った MCAW の解析では、よくアラインメントされた部位と文献で示唆されている結合認識に関与する糖鎖構造が一致することを確認できた。さらに、文献で示唆されていない構造部位についても MCAW は高いアラインメント表すことができた。MCAW の結果には、アラインメントの高い部位や低い部位に二分されず、中間の値を示す共通部位も見受けられたため、糖鎖結合タンパク質の結合認識は曖昧性をよく反映していることもわかった。

表１：マンノース特異的ジャカリン関連レクチンの糖鎖–レクチン相互作用実験を解析してまとめた文献(Nakamura-Tsuruta et al. 2008)と MCAW においてよくアラインメントされた部位の結果との比較。

レクチン名	解析結果
ジャックフルーツ、パラミツ Artocarpin	複合型糖鎖，Mana1-2Mana1-3Mana（弱い）Man α 1-6 枝の Galβ1-4GlcNAc
ブラックマルベリー MornigaM	Man α 1-6 枝の Galβ1-4GlcNAc
ヒロハヒルガオ Calsepa	バイセクティング GlcNAc の N 型糖鎖
キクイモ Heltuba	Mana1-2Mana1-3Mana1-6Man，Mana1-6Mana1-6Man，Mana1-2Mana1-3Man 枝の末端の a1-2Man
ソテツ CRLL	Mana1-2Mana1-2Mana1-3Man，Mana1-2Mana1-6Mana1-6Man，Mana1-2Mana1-3Mana1-6Man，末端の a1-2Man
タイワンバナナ BanLec	ハイマンノース型糖鎖のみ、Mana1-2Mana1-6Mana1-6Man 枝の末端の a1-2Man，Man7（100%一致）から Man9（約 20%）一致

LfDB の解析では、マンノース特異的ジャカリン関連レクチンについて糖鎖–レクチン相互作用実験を解析してまとめた文献(Nakamura-Tsuruta et al. 2008)と MCAW の結果を比較した。表１は、各レクチンの結合認識に関わっている部位を示す 。下線を引いた部位は、文献と MCAW の結果が一致した認識されると思われる糖鎖部分構造である。下線が引かれていない部位は MCAW の解析から得られた高いアラインメント結果である。このことから、MCAW の解析により、より広くレクチンの結合認識が発見できることが考えられる。

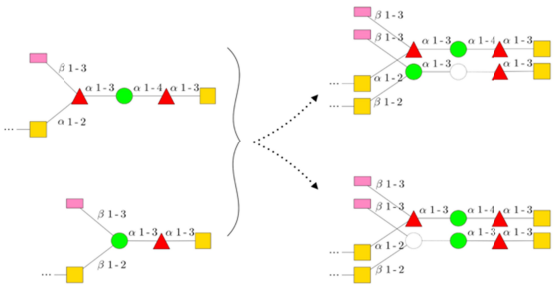

図２：分岐点にギャップを入れてアラインメントのできることを検討し、TED-MCAW を開発した。

なお、編集距離を用いたマルチプルアラインメントのツールは、TED-MCAW と名付け、現在 MCAW との性能の比較および検証を行っている。また、この検証結果をまとめた論文を現在執筆中である。さらに、TED-MCAW のアラインメント結果からスコア行列を作成し、このツールのアラインメント精度を今後さらに高める予定である。図２で示すように、TED-MCAW は分岐点に対するギャップ挿入を考慮していることから、糖鎖の確率モデルのためのより精度の高いステートモデルの決定に利用することができると考えられる。

このツールは現在 RINGS 上で公開されており、無償で利用することが可能となっている。また、このツールは研究者がデータ解析を容易にするための Web API も提供している。これにより、今後研究者が大規模な糖鎖の構造解析を行う際に、利用しやすい解析環境を提供することができると考えられる。

加えて、TED-MCAW は木編集距離をベースにしたアルゴリズムであるため、糖鎖構造同士の距離を直接計算することができる。この特性を利用し、今後糖鎖構造データベースにおける糖鎖構造の検索エンジンへと組み込むなどの応用ができることが期待できる。

５．主な発表論文等

〔雑誌論文〕（計 1 件）

① Masae HOSODA, Yukie AKUNE and Kiyoko F. AOKI-KINOSHITA. Development and application of an algorithm to compute weighted multiple glycan alignments 、Bioinformatics、査読有、Vol. 33、No. 9、2017、pp. 1317-1323.

〔学会発表〕（計 3 件）

① Masae HOSODA, Yukie AKUNE and Kiyoko F. AOKI-KINOSHITA. Lectin recognition pattern analysis using MCAW. Meeting Registration for: SFG & JSCR 2014 Joint Annual Meeting. 2014 年 11 月. Honolulu (USA).

② Masae HOSODA and Kiyoko F. AOKI-KINOSHITA. Analysis and development MCAW tool for elucidating recognized glycan patterns by glycan binding proteins. Warren Workshop VI 2016. 2016 年 8 月. 北海道大学（北海道札幌市）.

③ Masae HOSODA and Kiyoko F. AOKI-KINOSHITA. Development of a tool for extracting common glycan patterns recognized by avian influenza a virus. 2016 Society for Glycobiology Annual Meeting. 2016 年 11 月. New Orleans (USA).

〔図書〕（計 2 件）

① Kiyoko F. AOKI-KINOSHITA. Springer Japan. Glycoscience: Biology and Medicine. 2015. pp. 201-207.

② Kiyoko F. AOKI-KINOSHITA. Springer Japan. A Practical Guide to Using Glycomics Databases. 2017. pp. 299-334.

〔その他〕
ホームページ等
http://www.rings.t.soka.ac.jp/cgi-bin/tools/MCAW/mcaw_index.pl

６．研究組織

(1)研究代表者
木下　聖子（AOKI-KINOSHITA, Kiyoko）
創価大学・理工学部・教授
研究者番号：50440235

(4)研究協力者
細田　正恵（HOSODA, Masae）
高橋　悠志（TAKAHASHI, Yushi）