公募研究：生命システム情報

木確率モデルを用いた知識発見よりのタンパク質の糖鎖認識部位予測

●木下 聖子　◇西原 祥子
創価大学工学部生命情報工学科

＜研究の目的と進め方＞

本研究の目的はレクチンなどのタンパク質の糖鎖構造の認識機構を明らかにするための新しいデータマイニング技術の開発である。KEGG を含め、世界中に糖鎖構造情報に関するデータベースが増えつつある。また、糖鎖結合親和性を短時間で測定できる技術も発展してきており、糖鎖認識パターンを抽出するデータマイニング技術の糖鎖への応用が可能となってきている。しかし、現時点においては、糖鎖情報のための確率モデルの開発はまだ初期段階にある。研究代表者は、これまで Profile PSTMM と呼ぶ木構造中の子の順序を考慮する木確率モデルを開発し、糖鎖構造に応用してきたが、糖鎖の複雑な認識機構の本質をより良く理解するためには、更なる開発が必要である。本研究では、実際に糖鎖の認識部位予測に応用ことを目指して、これらの糖鎖確率モデルの拡張を行う。

＜ 2008 年度の研究の当初計画＞

本年度は Profile PSTMM と呼ぶプロファイルを抽出する確率モデルの計算上の精度の向上だけではなく、生物学的意義に関する確度についても改良を行う。このモデルを実用的に利用可能なものにするためには、糖鎖に特有な特徴を考慮する必要がある。まず、(1) 正確な比較を行うために、糖結合様式の情報をモデルに含める。これは現在含まれていない糖結合のアノマー配置情報や、結合に関与する炭素の番号に関する情報のことである。単純にノードのラベルに含めることも考えられるが、アノマーや結合に関与する炭素の番号を別のラベルとして扱うことも考えられるため、実験により精度を測りながら改良点を絞る予定である。(2) 更に、生化学的な類似性を考慮するために、この糖結合様式情報を組み込んだモデルのパラメータを学習するアルゴリズムを改良する。最も基本的なレベルでは、単糖はまず炭素の数によって分類できるため、ヘキソース（六炭糖）はペントース（五炭糖）よりも他のヘキソースとの類似性が高くなければならない。しかし、糖結合情報のモデルへの組み込みは単糖だけではなく、結合の生化学的類似性を考慮する必要もあるため、本研究代表者が以前開発した糖結合を含めた糖鎖スコア行列（置換行列）での計算を応用する。このアルゴリズムではまず、KCaM (KEGG Carbohydrate Matcher) と呼ぶ糖鎖の木構造アラインメントのアルゴリズムで、全ての糖鎖対のアラインメントを行う。アラインメントのスコアから、最も出現頻度の高い糖結合を抽出し、それぞれの糖結合対の確率の Log odds を計算する。従って、頻繁にアラインメントされる糖結合対は生化学的にも類似性があると考えられる。この手法はさらに解析する必要があるが、このようなスコア行列を利用して Profile PSTMM の学習アルゴリズムに結合情報を含めることができる。

また、本年度に Profile PSTMM の新しいモデルを検証するために、まず人工的に生成した糖鎖情報および実際の糖鎖情報の予測精度を測る。そのため、KEGG GLYCAN の糖鎖データベースの情報を主に使用し、実際の抽出されたモチーフを確認するために、Glycosciences.de や CFG (Consortium for Functional Glycomics) のデータベースも利用する予定である。

＜ 2008 年度の成果＞

以前開発した Profile PSTMM モデル のウェブツールを本年度新たに開発した。最も重要な問題点は入力された糖鎖構造に適応する State Model を確定することであった。Profile PSTMM のモデルは State Model の形に添ってプロファイルを出力する。そのため、入力された糖鎖構造に対して State Model の形が不適切であると、不正確なプロファイルを出力する可能性が生じる。この改善のために、初めに入力された糖鎖構造を用いて KCaM でのアラインメントを行うことにした。このアラインメントの結果から最大共通部分を取得し、State Model の形を確定することが出来る。この手法で RINGS (Resource for Informatics of Glycomes at Soka) のウェブツールの一つとして Profile PSTMM を利用可能にした。Profile PSTMM ツールでは、入力された糖鎖構造は KCF (KEGG Chemical Function) 形式で受け、自動的に最大共通部分木を抽出する。最大共通部分木から State Model の形を確定し、糖鎖のプロファイルを出力する。出力されたプロファイルは画像として表示するようにした。図１がその結果の例である。

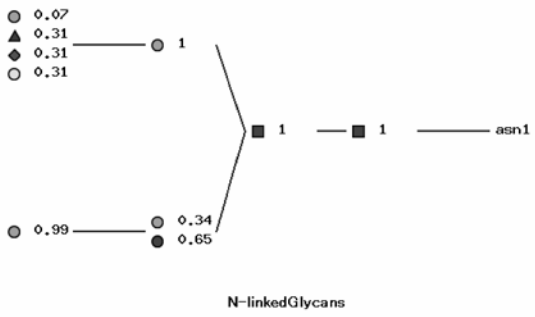

図１：Profile PSTMM のウェブツールが出力した *N* 型結合糖鎖のプロファイル

しかしながら、入力された糖鎖構造の数によって、最大共通部分木が小さくなる傾向が認められた。そのため、更に適切な State Model の確定法の研究を進め、糖鎖のマルチプルアラインメントからのブロックを抽出するアルゴリズムを開発した。糖鎖のブロックは、マルチプルアライメントから得たギャップのないモチーフ構造を示す。糖鎖ブロックを決定することによって State Model の構造を決められると考え、まずアミノ酸配列のマルチプルアライメントアルゴリズム CLUSTALW を基に、糖鎖構造のアルゴリズム MCAW (Multiple Carbohydrate Alignment

with Weights) を考えた。このアルゴリズムでは糖鎖の類似性から案内木 (guide tree) を作成し、有根木に変換する際に重み付けをする。そして、案内木に従って各糖鎖をマルチプルアライメントに追加して行き、重みをスコアの計算で用いることで、糖鎖のブロックを作成することができる。この糖鎖ブロックから適切な State Model を構築できると考えられた。

MCAW アルゴリズムはまず、KCaM の Exact Match アルゴリズムを用いて全糖鎖の類似性を比較し、KCaM の類似スコアを得る。このスコアは最大 100% となり、二つの糖鎖構造の類似度を指す値である。クラスタリングを行うために類似スコアを距離スコアに変換し、行列に置き換える。そして、Fitch-Margoliash 法で全糖鎖のクラスタリングを行い、案内木を作成し、各糖鎖構造の重みを得る。この案内木を基に、糖鎖構造を次第にマルチプルアライメントに追加して行く。

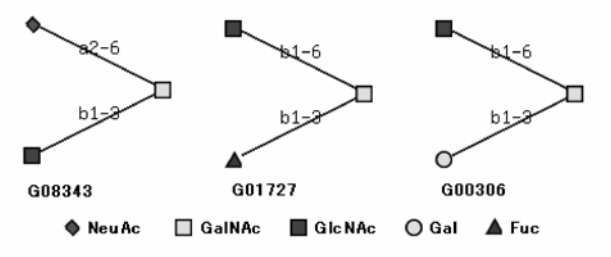

図２：糖鎖のマルチプルアライメントの例

例えば、図２のような糖鎖構造をアライメントする場合、上の位置にある糖のスコアは次の式を用いて計算する：

$$S = (Q(NeuAc\ \ [a2-6], GlcNAc\ \ [b1-6]) * w1 * w2 + Q(NeuAc\ \ [a2-6], GlcNAc\ \ [b1-6]) * w1 * w3 + Q(GlcNAc\ \ [b1-6], GlcNAc\ \ [b1-6]) * w2 * w3) / 3$$

ここで、*w1, w2, w3* は G08343, G01727, G00306 の糖鎖のそれぞれの重みである。このスコアが KCaM のダイナミックプログラミングアルゴリズムに用いられ、全体のマルチプルアライメントを計算する。

糖鎖構造のマルチプルアライメントを行うための形式も決める必要があったため、PKCF (Profile KCF) と呼ぶ新しい糖鎖アライメント形式も開発した。そして、この形式を扱うことができる MCAW のプログラムを開発した。

＜国内外での成果の位置づけ＞

現在に至るまで、木構造のマルチプルアライメントは主に系統樹のアライメントのために開発されてきた。そのため、本研究において、木のマルチプルアライメントのアルゴリズムを新たに開発する必要があった。MCAW は糖鎖構造のマルチプルアライメントの初めてのアルゴリズムである。この結果から、今後 Profile PSTMM の State Model を効率よく決定することができた。また、糖鎖のブロックから糖鎖「ファミリー」のモチーフの抽出などにも応用できると考えられる。例えば、sialyl-Lewis X は癌部で発現する腫瘍マーカーとして知られている。糖鎖は様々な生物学的プロセスにおいて重要な役割を果たすと報告されているが、一方、sialyl-Lewis X のような生物学的な機能と結びついた糖鎖モチーフの発見は遅れているため既知のモチーフ数はまだ少ない。本研究で提示する糖鎖ブロックの情報より、新たな糖鎖モチーフが発見できるものと考えられる。このような糖鎖モチーフ解析はバイオインフォマティクスの分野ではほとんど報告例はなく、本研究の成果は、糖鎖の医療や創薬への貢献を加速すると期待できる。

＜達成できなかったこと、予想外の困難、その理由＞

当初は State Model の確定法の研究を計画に含めていなかったため、単糖と結合の情報をモデルに含める研究を進めることができなかった。

＜今後の課題＞

今後は MCAW を用いて糖鎖ブロックを作成し、今年度の計画であった単糖と結合情報の調査および糖鎖スコア行列の作成に取り組む。同時に糖鎖スコア行列を Profile PSTMM に組み込むアルゴリズムを開発する。この新しいアルゴリズムを検証するため、CFG の糖鎖親和性データを利用して予測精度を計る。最終的にタンパク質が認識する糖鎖構造を予測し、生物学的な評価を行う。

＜成果公表リスト＞

今年度に関しては該当なし。