公募研究：2008 ～ 2009 年度

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、世界中に糖鎖構造データベースの情報が増えつつある。また、糖鎖結合親和性の高速な測定技術も発展しているため、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いる。現時点において、糖鎖情報のための確率モデルの開発はまだ初期段階である。研究代表者は、これまで ProfilePSTMM 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良く理解するために更なる開発が必要である。本研究では、実用的な糖鎖の認識部位予測に応用するため、これらの糖鎖確率モデルの拡張を行う。

＜研究開始時の研究計画＞

開始時の研究計画において、は ProfilePSTMM 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、生物学的意義についての精度も改良することであった。このモデルを実用的に利用できるためには糖鎖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、ヒドロキシル基のことである。単純にノードのラベルに含めることも考えられたが、アノマーやヒドロキシル基を別のラベルとして扱うことも考えられたため、実験により精度を測りながら改良点を絞る予定であった。さらに、(2) 生化学的な類似性を考慮するために、この糖結合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計画であった。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）での計算を応用できる。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 score 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 ProfilePSTMM 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、ProfilePSTMM 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計画であった。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った。

＜研究期間の成果＞

以前開発した ProfilePSTMM モデル [1,2] のウェブツールを開発してきた。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。ProfilePSTMM 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 [2] 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 ProfilePSTMM を利用可能にした。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。

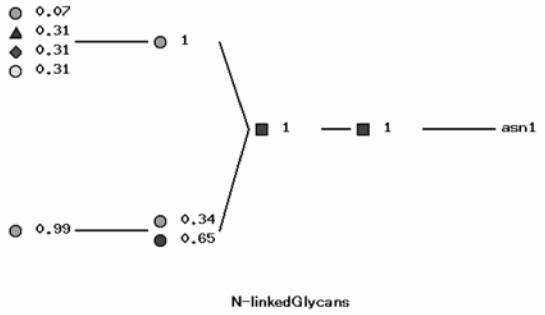

図１：ProfilePSTMM のウェブツールが出力した N 型結合糖鎖のプロファイル。

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見られた。そのため、State model の形を確定するために、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) を考案した。

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% とし、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。

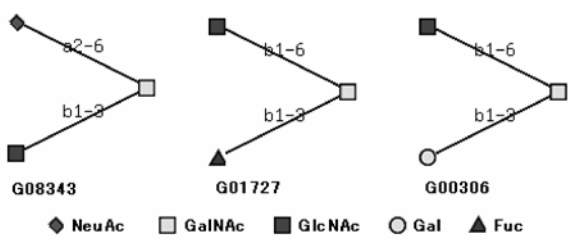

図２：糖鎖のマルチプルアライメントの例

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（結合を含む）糖のスコアは次の式を用いて計算する：

*S=(Q(NeuA [a2-6], GlcNAc [b1-6])*w1*w2+*
*Q(NeuAc [a2-6], GlcNAc [b1-6])*w1*w3+*
*Q(GlcNAc [b1-6], GlcNAc [b1-6])*w2w3)/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。このスコアを KCaM のダイナミックプログラミングアルゴリズムに応用し、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、糖鎖のマルチプルアラインメント形式も決める必要があり、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。

PKCF では、入力された糖鎖構造がアライメントされた順に従って、1 つのポジションにおける各ノード情報が表示されている。たとえば、KEGG GLYCAN で定義される G 番号、G02183 と G05700 のアライメントの結果で得られた PKCF は図３の様になる。NODE 5 の行で表示される “5　1=GlcNAc　2=GlcNAc” からは、5 のポジションにある単糖は 1(G02183) が GlcNAc、また 2(G05700) が GlcNAc であることが分かる。さらに EDGE 情報は EDGE　4 と 5 の行で示され、“4　5-1:b1　4-1:2” の行から NODE　5-1 の GlcNAc は NODE　4-1 の Man と β 1-2 結合していることが、また “5　6-1:b1　5-1:4” から NODE　6-1 の Gal は NODE　5-1 の GlcNAc に結合していることが示されている。

また、本研究では「ノード間に生じるノードの挿入 / 欠失の状態」を「ギャップ (Gap)」と定義し、糖鎖構造の還元末端である根 (ルート) と非還元末端である葉 (リーフ) に生じるノードの挿入 / 欠失の状態」を「ミッシング (Missing)」と定義し、それ以外の一般のノードを「単糖 (Residue)」と定義し、アライメントの際にギャップあるいはミッシングが生じた場合には、これを PKCF に反映させるようにした。具体的には、ギャップは「－」（マイナス）でミッシングは「０」（ゼロ）として表示するようにした。

＜国内外での成果の位置づけ＞

ProfilePSTMM のウェブツールは世界的に初めての糖鎖プロファイルのマイニングツールである。そのため、国内外において反響が大きい。レクチンの研究のみならず、糖転移酵素の認識機構、糖鎖同士の結合機構への応用も考えられる。他のモデルでは検出することのできないパターンをプロファイルとして抽出できるため、このツールはユニークであり、様々な糖鎖の研究に応用可能である。

```
ENTRY       G02183-G05700          GlycanProfile
NODE  21
      1     1=GlcNAc   2=GlcNAc       0    0
      2     1=GlcNAc   2=GlcNAc      -8    -4
      3     1=Man   2=Man       -16    -7
      4     1=Man   2=Man       -24    -4
      5     1=GlcNAc   2=GlcNAc      -32    -5
      6     1=Gal   2=Gal       -40    -5
      7     1=GlcNAc   2=-       -48    -5
      8     1=Gal   2=-       -56    -5
      9     1=Neu5Ac   2=Neu5Ac      -64    -5
      10     1=GlcNAc   2=GlcNAc      -32    -3
      11     1=Gal   2=Gal       -40    -3
      12     1=GlcNAc   2=-       -48    -3
      13     1=Gal   2=-       -56    -3
      14     1=Neu5Ac   2=Neu5Ac      -64    -3
      15     1=Man   2=Man       -24    -10
      16     1=GlcNAc   2=GlcNAc      -32    -11
      17    1=Gal   2=Gal       -40    -11
      18     1=Neu5Ac   2=Neu5Ac      -48    -11
      19     1=LFuc   2=0       -8    4
      20     1=GlcNAc   2=0       -32    -9
      21     1=Gal   2=0       -40    -9
EDGE   20
      1     2-1:b1     1-1:4
      1     2-2:b1     1-2:4
      2     3-1:b1     2-1:4
      2     3-2:b1     2-2:4
      3     4-1:a1     3-1:6
      3     4-2:a1     3-2:6
      4     5-1:b1     4-1:2
      4     5-2:b1     4-2:2
      5     6-1:b1     5-1:4
      5     6-2:b1     5-2:4
      6     7-1:b1     6-1:3
      6     7-2     6-2
      7     8-1:b1     7-1:4
      7     8-2     7-2
      8     9-1:a2     8-1:3
      8     9-2:a2     8-2:6
      9     10-1:b1     4-1:6
      9     10-2:b1     4-2:6
      10     11-1:b1     10-1:4
      10     11-2:b1     10-2:4
      11     12-1:b1     11-1:3
      11     12-2     11-2
      12     13-1:b1     12-1:4
      12     13-2     12-2
      13     14-1:a2     13-1:3
      13     14-2:a2     13-2:3
      14     15-1:a1     3-1:3
      14     15-2:a1     3-2:3
      15     16-1:b1     15-1:2
      15     16-2:b1     15-2:2
      16     17-1:b1     16-1:4
      16     17-2:b1     16-2:4
      17     18-1:a2     17-1:3
      17     18-2:a2     17-2:3
      18     19-1:a1     1-1:6
      18     19-2     1-2
      19     20-1:b1     15-1:4
      19     20-2     15-2
      20     21-1:b1     20-1:4
      20     21-2     20-2
///
```

図３：G02183 と G05700 のアライメントの結果で得られた PKCF

＜達成できなかったこと、予想外の困難、その理由＞

最も予想していなかったポイントとして、State Model の改善のための時間であった。2 年間の計画の内、1 年目に完成させる予定が 2 年目に延びてしまった上、複雑な糖鎖構造に対応させるために時間が通やしてしまった。従って、2 つの糖鎖構造のアラインメントを行うことができたが、アラインメントとアラインメントのマルチプルアラインメントはいまだに完成していない。その理由として、糖鎖構造の結合情報及び分岐する形によって生じた問題で、デバッグするための時間が予想以上、大幅に時間がかかってしまったことにある。今年度末までにはこれを完成させ、計画の引き続きを来年度に、実際のレクチンのデータを用いて検証を行う予定である。

＜今後の課題、展望＞

レクチンのデータを用いて検証を行う計画に加え、糖鎖のスコア行列を考慮する課題も残っている。現在、新たにスコア行列のアルゴリズム及びツールを開発中である。このツールの一部として MCAW のプロファイル結果が必要とする。従って、MCAW のアラインメント対アラインメントのプログラムを今年度末に完成させ、これをスコア行列のツールで利用できるようにする。また、その結果を再度 MCAW で利用できるようにするため、MCAW のプログラム改善が続く予定である。

また、近年、PSTMM の改善されたモデルとして Ordered Tree Markov Model (OTMM) を開発した。このモデルは以前の PSTMM と同様な精度を得ながら overfitting 問題を解決できた。従って、現在の ProfilePSTMM モデルを ProfileOTMM モデルにも改良し、開発する予定である。

今後の展望として、今まではデータマイニングを使用し、レクチンや糖転移酵素の認識部位予測が行われなかった。一つの理由として、糖鎖の複雑な合成機構や構造が挙げられた。しかし、本研究はこの壁をはじめて乗り越えられるようにした。

＜研究期間の全成果公表リスト＞

1）論文／プロシーディング

なし

2）学会発表

池田秀一、木下聖子、西原祥子Development of a tool for mining glycan tree structures、BMB2008、2008年12月9日、神戸

海谷咲希子、木下聖子Probabilistic Tree model; Improvement of ProfilePSTMM、JSBi、2008年12月9日、神戸

池田秀一、木下聖子、The profile PSTMM tool for mining glycan tree structures、JSBi、2008年12月15日、大阪

3）図書

なし

4）データベース／ソフトウェア

http://rings.t.soka.ac.jp/profile-training.html

5）研究成果による産業財産権の出願・取得状況

なし