OTMM を用いた糖鎖の解析手法

指導教員　**篠宮 紀彦**
1958207　**戸塚 健人**

1　はじめに

糖鎖は, DNA・タンパク質に次ぐ「第 3 の生命鎖」と呼ばれており, 細胞表面のタンパク質や脂質に結合し, 多くの生命現象に関与している. 糖鎖とは, グルコースやガラクトースなどの単糖が繋がって生成される「糖の鎖」であり, 多種多様な単糖が分岐をして繋がっていくため, 複雑な構造をとる. また, その機能は糖鎖単体では決まらず, タンパク質や脂質などの「糖鎖が結合する相手」や時期や環境などの「状況」に応じて大きく異なる. さらに, 糖鎖はレクチンと総称されるタンパク質に特異的に認識されることがわかっており, レクチンが糖鎖の多様な働きに関与していると考えられている. 以上のように, その複雑な構造と多様な反応性から糖鎖には未だに多くの謎が存在し, 糖鎖の本質的な機能の解明やその利用に期待が高まっている. このようなことから, 本研究では計算機科学の知識と生物学の知識を融合し, これまでの研究では見出せなかった糖鎖に潜む法則性を発見する. そして, そのような発見が「生命現象の本質的な解明」, 「糖鎖を目印とした病気の発見」, 「創薬・ワクチン開発」などで役立てられることを期待する.

2　研究の目的

糖鎖とは, 単糖が鎖状に繋がった物質である. 哺乳類から植物に至るまであらゆる生物種に存在し, 細胞表面の脂質やタンパク質に結合している. また, 糖鎖の細胞への付加率は常に一定ではなく, 結合する際の構造も常に同じではない [1]. このような「糖鎖の不均一さ」によって, 糖鎖解析は困難になっている.

糖鎖を認識する糖結合タンパク質の総称が「レクチン」である. レクチンは, 多種多様な糖鎖から特定の糖鎖を同定することができる [2]. 例えば, レクチンの一種である「ガレクチン」はガラクトースを含む糖鎖を認識し結合する. このように, 一部のレクチンは様々な糖鎖を幅広く認識する. したがって, レクチンが認識する糖鎖を理解するためには, 認識される糖鎖構造の共通パターンを調べる必要がある.

本研究の目的は, 「糖鎖構造の共通パターン」を正確に取得できる確率モデルを開発することである. 具体的には, 特定の種類に絞った糖鎖構造を学習し, その共通パターンを取得する. アルゴリズムの設計手法としては, 動的計画法と機械学習が用いられている [3].

3　先行研究

前述の通り, 糖鎖は複雑な構造を取るので, グラフ理論の「木」を用いてその構造を表す. 木とは「連結かつ閉路がないグラフ」のことである. 先行研究では「隠れマルコフモデル (HMM)」をベースにした機械学習を用いて, 糖鎖データからその構造の共通パターンを状態遷移図として取得している [4, 5]. ここで, 「糖鎖データ」とは, 糖鎖構造をテキストで表現したデータを意味する. このデータを右から左へ読み取ることで, 根元から葉に向かってノード (単糖) の繋がり方を取得することができる. すなわち, 糖鎖構造を取得することができる. 以降, 糖鎖構造のデータを「糖鎖データ」と呼ぶ.

[4] の研究では, $PSTMM(Probabilistic\ Sibling-dependent\ Tree\ Markov\ Model)$ と呼ばれる確率モデルを用いて糖鎖構造の解析を行っている. $PSTMM$ は, 糖鎖構造を詳細に表現できる一方, 親子間や兄弟間でノードが依存し合っているため, 過学習が発生しやすい.

[5] の研究では, $OTMM(Ordered\ Tree\ Markov\ Model)$ と呼ばれる確率モデルを用いて糖鎖構造の解析を行っている. $OTMM$ は, $PSTMM$ と同レベルの正確性を維持しながら, 過学習の問題を改善することができている. また, $PSTMM$ に比べ大幅に計算時間が短縮している.

4　OTMM による糖鎖データの学習

糖鎖には未だに不明な点が多く, その解析は困難を極める. したがって, 本研究では, 学習する糖鎖の種類を「N-結合型糖鎖」に絞る. また, $OTMM$ に新しいタイプの糖鎖データを入力することで, 既出の論文とは異なる新しい解析結果を得ることを目指す. 具体的には, 既存の研究では「単糖」のみの糖鎖データを学習していたが, 本研究では「単糖」と「結合様式」を組み合わせた糖鎖データを学習する.

以下, $OTMM$ は 5 つの要素で定義され, 観測可能なラベルと観測できない状態が存在するのが特徴である.

OTMM

$S = \{s_1, \cdots, s_{|S|}\}$:(観測できない) 状態の集合

$\Sigma = \{\sigma_1, \cdots, \sigma_{|\Sigma|}\}$:(観測可能な) ラベルの集合

$\Pi = \{\ \pi\ (l)\}$: 初期状態確率分布

例：$\pi\ (l)$ はルート (根) の状態が s_l である確率

a : 状態遷移確率分布

$\alpha[q, m]$: 状態 q の親から状態 m の子へ遷移する確率

$\beta[l, m]$: 状態 l の兄から状態 m の弟へ遷移する確率

b : ラベル出力確率分布

例：b[l, σ_h] はノードの状態が s_l の時, ラベル σ_h が出力される確率

本研究では, 観測可能な糖鎖データから隠れた糖鎖構造の共通パターンを見つけ出すことが目的である. また, 学習では $EM(Expectation - Maximization)$ アルゴリズム [5] を用いる. このアルゴリズムでは, 入力された糖鎖データから, 初期状態確率, 状態遷移確率, ラベル出力確率 (3 種類の確率分布) を求める. また, 観測可能な糖鎖データからその構造の特徴を取得する際は $Viterbi$ アルゴリズム [5] を用いる.

5　実験結果

本研究では, $OTMM$ に「単糖のみ」の糖鎖データと「単糖 + 結合様式」の糖鎖データをそれぞれ入力し, 比較を行った. 図 1 に処理時間の比較を示す.

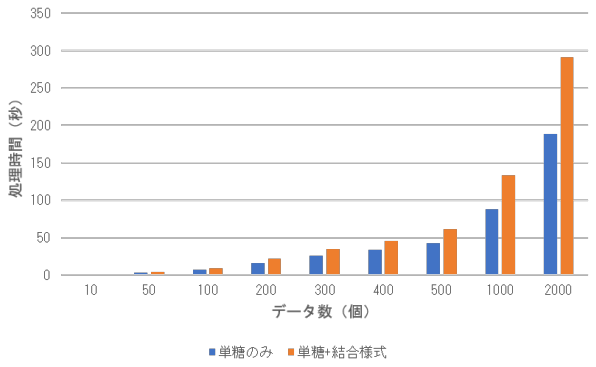

図 1　OTMM における 1 エポック当たりの処理時間の比較

「単糖のみ」の糖鎖データを用いた既存の研究と比べて, 「単糖 + 結合様式」の糖鎖データを用いた本研究は, 処理時間が長くなった. これは, 「単糖 + 結合様式」の糖鎖データは, 必ず「単糖のみ」の糖鎖データ以上の情報量を持つためであると考えられる.

続いて, 1000 個の「単糖 + 結合様式」の糖鎖データを学習したモデルを用いて, 糖鎖構造の解析 $(Parsing)$ を行った. 図 2 に解析結果の一部を示す.

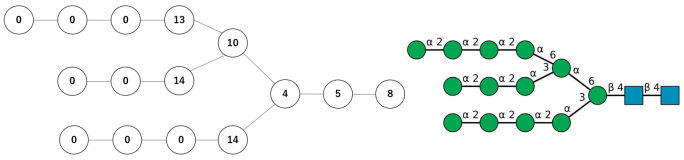

図 2　学習結果 (左) と解析に用いた糖鎖（右）

右の糖鎖構造において, 単糖に相当するのが頂点であり, 結合様式に相当するのが辺の文字である. 左の解析結果から, 状態を表す番号が, 単糖とその結合様式に応じて, 各ノードに割り振られていることがわかる. このような微細な構造の違いは, 「単糖のみ」の学習で捉えることができない. したがって, 「単糖 + 結合様式」の学習によって, より正確に糖鎖構造の特徴を得ることができた.

6　おわりに

本研究では, $OTMM$ に新しい形式の糖鎖データを入力することで, 糖鎖構造の共通パターンをより正確に取得することを目指す. 実験結果から, 単糖と結合様式を組み合わせたデータを学習することによって, より正確に糖鎖構造の特徴を得ることができるとわかった. しかしながら, 処理時間は「単糖のみ」の学習の方が短くなるため, 正確性と効率性がトレードオフの関係にあることもわかった. 本研究が完成することで, 糖鎖インフォマティクスや糖鎖の基礎研究を進展させることができ, 糖鎖科学の発展に寄与できると確信している. 今後は, $OTMM$ を $profileOTMM$ というモデルに改良する. これによって, $OTMM$ では困難であった「容易に糖鎖構造の共通パターンを取得すること」が可能になる. さらに, レクチンの種類別に糖鎖データを学習することで, 各レクチンが認識する糖鎖構造を取得することも目指す.

参考文献

[1] 平林淳, 糖鎖とレクチン, 日刊工業新聞社, 2016.

[2] 長江 他, 糖鎖の多様性に対応するレクチンの認識システムとシグナリング, 生化学第 90 巻第 5 号, 2018, pp.651–663.

[3] 渋谷 他, (2007). バイオインフォマティクスのためのアルゴリズム入門, 共立出版.

[4] N.Ueda, et al. (2005). A probabilistic model for mining labeled ordered trees: capturing patterns in carbohydrate sugar chains, IEEE transactions on knowledge and data engineering, vol.17, no.8.

[5] K.Hashimoto, et al. (2008). A new efficient probabilistic model for mining labeled ordered trees applied to glycobiology, ACM transactions on knowledge discovery from data, vol.2, no.1.