Vol.2013-BIO-34 No.30
2013/6/28

機械学習を用いたマススペクトルデータからの糖鎖構造推定法の開発

雲崎 翔太郎[1]　佐藤 健吾[1,a)]　榊原 康文[1]

概要：近年，タンパク質を介しさまざまな調節を受けて生合成される糖鎖を扱うグライコミクス研究が世界中で盛んに行われ急速な技術発展を遂げている．現在では多糖構造の同定においてタンデムマススペクトロメトリー (MS/MS) が主要なツールとなっている．しかしマススペクトルデータからその糖鎖構造を解析して決定する計算手法の開発は少なく，汎用的なソフトウェアは少ない．先行研究において，構造を決定する上で有用な開裂イオンが持つ情報を用い，動的計画法を利用することで，マススペクトルデータから糖鎖構造を決定する手法が提案された．しかし，スコア関数が最適化されていない，複雑な構造が扱えない，スペクトルデータ中のノイズに弱いなどの問題点があった．そこで本研究では，マススペクトルを生成する機械のノイズや偏りを統計的機械学習手法によって学習することによって，マススペクトルデータから糖鎖構造をより正確に解析できるアルゴリズムを開発した．

1. 背景

近年，タンパク質を介しさまざまな調節を受けて生合成される糖鎖を扱うグライコミクス研究が世界中で盛んに行われ急速な技術発展を遂げている．

糖鎖は，最小単位である単糖が繋がってできている．その構造は単糖の種類，単糖と単糖の繋がり方を表す linkage 情報（グリコシド結合の結合位置）によって表現され，核酸やアミノ酸の配列と異なり，図 1 (左) のように木構造を用いてモデル化される．現在では糖鎖構造の同定においてタンデムマススペクトロメトリー (MS/MS) が主要なツールとなっている．糖鎖試料を断片化し，MS/MS によって観測された質量電荷比（m/z 値）を横軸，相対強度を縦軸とするグラフを描くと，各ピークは糖鎖断片に対応する (図 2)．糖鎖の断片化は，単糖と単糖の間で切断されて起こる場合（基本ピーク）の他に，図 1（右）のように単糖の環状構造内部で開裂して起こる場合（派生ピーク）があり，この点がアミノ酸の断片化と大きく異なる．このような派生ピークは，開裂のタイプ，単糖の種類等の組み合わせを考えると 528 通りある．そこで，先行研究 GLYCH [2] では，開裂によって生じうる断片の質量から基本ピークの m/z 値を逆算することによって，観測されたすべての派生ピークを基本ピークに変換し，各々の基本ピークをサポートしている派生ピークの数（サポートピーク数）を計算する．そして，サポートピーク数を最大とするような基本ピークを持つ糖鎖の木構造を動的計画法によって求める．

しかし，サポートピーク数に基づくスコア関数が簡易的であり，複雑な構造が扱えない，スペクトルデータ中のノイズに弱いなどの問題点がある．そこで，本研究では統計的機械学習を用いて GLYCH を改良し，マススペクトルを生成する機械のノイズや偏りを学習することによって，マススペクトルデータをより正確に解析できるアルゴリズムを開発した．

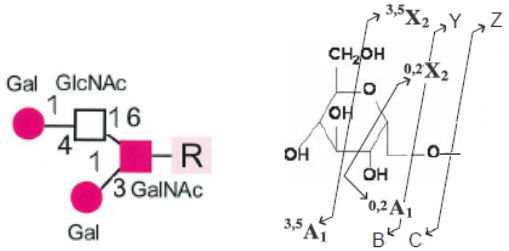

図 1　(左) 木構造による糖鎖構造のモデル化　(右) 糖鎖の開裂

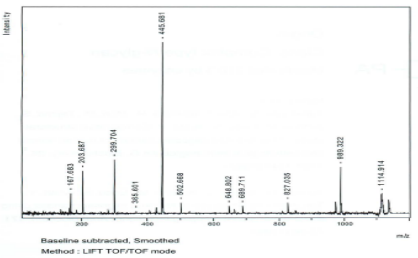

図 2　マススペクトルデータ

1　慶應義塾大学理工学部生命情報学科
a)　satoken@bio.keio.ac.jp

ⓒ 2013 Information Processing Society of Japan

Vol.2013-BIO-34 No.30
2013/6/28

2. 手法

本研究では，先行研究 GLYCH に新たなスコア関数を加え，さらにそれらを機械学習で最適化することによって精度の向上を目指す．糖鎖の解析においては，解析手法，実験手法により偏りが生じると考えられるため，統計的機械学習によってこの偏りを学習できると考えた．GLYCH においては，単純にサポートピーク数をスコアとしていたのに対して，本研究では，開裂の起こりやすさや単糖と単糖の結合しやすさを表すスコアを導入する．観測された N 個のピーク M_i $(i = 1, \ldots, N)$ が与えられた時，以下の再帰式に基づいて，与えた糖鎖全体の質量 $M_{parent}(= M_N)$ のスコア $V(M_{parent}, r, b)$ が最大となるような構造を動的計画法によって求める．

$$V(m,r,b) = \sum_i \sum_c I(m = M_i + \delta M_{rc}) E(r,b,c) + \max_{m_1 \le m_2 < m} \begin{cases} 0 & (\text{if } m = \text{mass}(r)) \\ T(r, r_1, b_1) + V(m_1, r_1, b_1) & \\ \qquad (\text{if } m = \text{mass}(r) + m_1) & \\ B(r, r_1, b_1, r_2, b_2) & \\ \qquad + V(m_1, r_1, b_1) + V(m_2, r_2, b_2) & \\ (\text{if } m = \text{mass}(r) + m_1 + m_2,\ b_1 \neq b_2) & \end{cases} \quad (1)$$

ここで，m は基本ピークの m/z 値，r は単糖，b はグリコシド結合の結合位置である．$I(cond)$ は，条件 $cond$ が真の時 1，偽の時 0 を返す識別関数である．$\text{mass}(r)$ は単糖 r の質量を表す．$E(r,b,c)$ は，単糖 r が結合位置 b で結合し，開裂 c が起こる時のスコアである．δM_{rc} は，単糖 r で開裂 c が起きた時の断片の質量である．$T(r_i, r_j, b_j)$ は単糖 r_i が r_j と結合位置 b_j で結合する時のスコア，$B(r_i, r_j, b_j, r_k, b_k)$ は単糖 r_i で分岐が起こり，r_j と結合位置 b_j で，r_k と結合位置 b_k で結合する時のスコアである．

開裂のしやすさを表すスコア関数 $E(r,b,c)$，単糖の結合しやすさを表すスコア関数 $T(r_i, r_j, b_j)$，単糖の分岐しやすさを表すスコア関数 $(r_i, r_j, b_j, r_k, b_k)$ は，MIRA [1] と構造化 SVM [3] を用いて訓練データから学習する．これらの学習法は，誤った部分にペナルティを課すことによってパラメータを更新することを繰り返し，次第に間違った構造を予測しにくくする手法である．

3. 結果と考察

訓練データとして，[4] から 3 分岐の糖鎖構造を除いた 100 個の糖鎖のスペクトルデータを抽出して用い，10 分割の交差検定で精度を評価した．予測精度は，予測構造におけるグリコシド結合の Positive Predictive Value (PPV) と Sensitivity で評価した．すなわち，各々のグリコシド結合について，両端における単糖の種類と結合位置が等しい時に正解とした．学習しない場合（= GLYCH）と構造化 SVM，MIRA で学習した場合の PPV と Sensitivity を図 3，4 に示す．いずれの学習アルゴリズムにおいても，予測精度の向上を確認することができた．

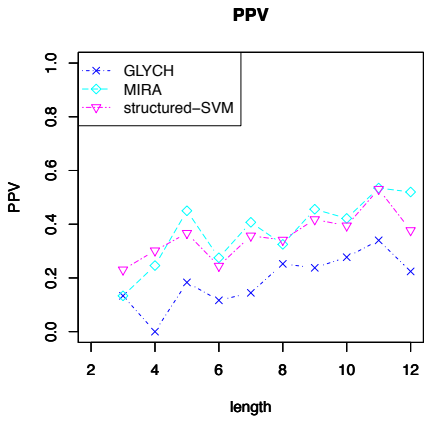

図 3　予測精度 (PPV)

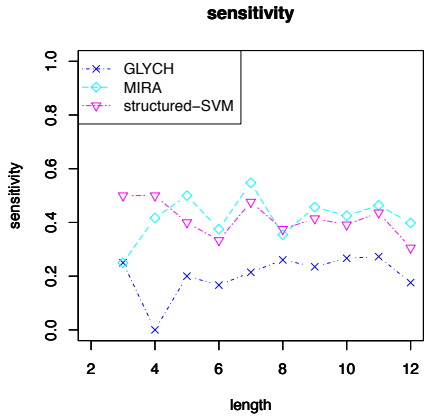

図 4　予測精度 (Sensitivity)

4. 結論

MS/MS で得られたスペクトルデータより糖鎖構造を推定する既存手法 GLYCH のスコア関数の改良し，さらにそれらを統計的機械学習を用いて訓練データから学習することによって予測精度の向上を実現した．今後は，スコア関数の改良を検討する．また，実測データのさらなる収集を行い，予測精度の向上を目指す．

参考文献

[1] Crammer, K. and Singer, Y.: Ultraconservative online algorithms for multiclass problems, *Journal of Machine Learning Research*, Vol. 3, pp. 951–991 (2003).
[2] Tang, H., Mechref, Y. and Novotny, M. V.: Automated interpretation of MS/MS spectra of oligosaccharides, *Bioinformatics*, Vol. 21 Suppl 1, pp. i431–439 (2005).
[3] Tsochantaridis, I., Joachims, T., Hofmann, T. and Altun, Y.: Large margin methods for structured and interdependent output variables, *Journal of Machine Learning Research*, Vol. 6, pp. 1453–1484 (2005).
[4] 黒河内政樹，広瀬和子：糖鎖の MALDI-TOFMS スペクトルデータブック，三共出版 (2007).

ⓒ 2013 Information Processing Society of Japan