G-011

確率モデルに基づく質量分析におけるマススペクトルの解析

Analysis of mass spectrum in mass spectrometry based on probability model

橘 勇人† 高橋 篤‡ 錦織 充広‡ 大星 直樹†
Isato Tachibana Atsushi Takahashi Mitsuhiro Nishigori Naoki Oboshi

1. 背景

質量分析（MS）とは分子をイオン化し, その 質量電荷比（m/z）を検出してイオンの質量を測定するものである。近年、質量分析器の精度向上やデータ解析技術の進歩により、タンパク質の網羅的な解析（プロテオーム解析）が大きく発展を遂げてきたが、今もなお発展段階にある。プロテオーム解析データから、新たなバイオマーカーや創薬標的の発見などが期待されている[1]。MS で広く用いられる液体クロマトグラフィータンデム質量分析法(LC/MS/MS)より得られた MS/MS スペクトルからタンパク質同定を行うフローは次の図 1 の通りである。

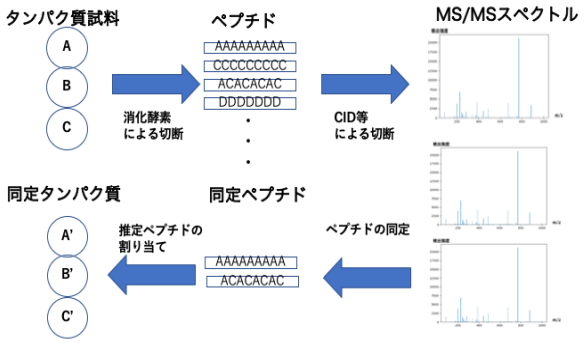


図 1　MS/MS によるタンパク質同定のフロー

1. タンパク質試料を消化酵素によりペプチドに切断後、LC により分離する
2. LC に連結した質量分析器で得られた MS スペクトルよりプリカーサーイオン（ペプチド）を選択
3. 衝突誘起解離(CID)によりイオンを解離させ、フラグメントイオンを検出する
4. 計測した MS/MS スペクトル情報からペプチド配列の同定を行う
5. アミノ酸配列データベース（DB）と照合し、同定したペプチド配列群がどのタンパク質の配列に割り当てられるかによりタンパク質を推定する。

特に、生体組織や体液試料を用いたプロテオーム解析では、得られるマススペクトルは数十万にも及ぶ。質量分析において、大量のマススペクトルからペプチド配列をより高精度に同定することは試料に含まれるタンパク質を網羅的に解析するための重要課題である。これら全てのマススペクトルを専門家が手作業で一つ一つ精査していくことは現実的ではない。現在、タンパク質同定の問題点として、膨大なマススペクトルの内、決して少なくない数のマススペクトルがペプチド配列を同定する際にスペクトルに十分なピークがないなどの理由で破棄されてしまっていることが挙げられる。そこで本研究では、Ionsearch[2]などの既存の手法の MS/MS のピークとアミノ酸配列 DB の仮想的な理論値と比較によりペプチドの同定を行う手法に対して、本研究では MS/MS スペクトルが混合正規分布による確率モデルから生成されたと仮定し、そのモデルを変分ベイズにより推定する。このモデルを利用しデータベースのアミノ酸配列の m/z の理論値に対してスコアリングすることでペプチドを同定する手法を提案する。

2. 提案手法

2.1 実験データと検索データベース

本研究で使用したデータはマウスの心臓組織を解析したもので、通常マウスと心不全マウスを比較し、どのようなタンパク質に変動があるかを調べる目的のものである。

また、そのタンパク質を検索するデータベースとしては Swiss-Prot[3]と呼ばれるタンパク質のアミノ酸配列知識ベースを利用し、理論上得られるマススペクトルのピークの m/z のテーブルを作成し、マススペクトルからペプチドの配列を同定する際に利用する。CID によって切断され、マススペクトルにピークとして検出されるペプチドのフラグメントイオンのテーブルは次の図 2 の通りである。

アミノ酸	b ion (m/z)	y ion (m/z)
I	114.0913	830.4519
H	251.1503	717.3679
F	398.2187	580.3089
G	455.2401	433.2405
A	526.2772	376.2192
T	627.3249	305.1819
G	684.3436	204.1343
K	812.4413	147.1128

図 2　アミノ酸配列(IHFGATGK)の
フラグメントイオンテーブル

このテーブルはペプチド配列[IHFGATGK]が CID によって切断およびイオン化された際の b-イオン(切断の N 端側で発生するイオン)と y-イオン（切断の C 端側で発生するイオン）の m/z の値を示しており、これらは理論上、MS/MS スペクトル上のピークとして出現する。この他にも脱 NH_3 やリン酸化などの翻訳後修飾により、m/z の異なるピークが検出される可能性があるが、本研究では考慮していない。

2.2 変分ベイズによる混合ガウスモデルの推定(VBGMM)を用いた同定

変分ベイズは確率モデルのパラメータの事後分布を求める手法の一つである[4]。これを利用し、マススペクトルが混合ガウス分布に従って生成されると仮定し、無作為に抽出

† 近畿大学院総合理工学部　Kindai University Graduate School of Science of Engineering Faculty of Science and Engineering
‡ 国立循環器病研究センター　National Cerebral and Cardiovascular Center

したマススペクトルの合がどのようなパラメータの混合ガウスモデルに従って生成されたかを推定する。推定した混合ガウスモデルを利用し、データベース内のペプチドの理論上の m/z ピークを標本とし対数尤度の平均を求める。この値からデータベースのペプチドが推定したモデルにどの程度一致しているかスコアとして算出し、利用することでペプチド配列の同定を行う。(図3)。このスコアは大きければ大きいほど推定した混合ガウス分布への当てはまりが大きくなり、利用したマススペクトルに含まれている確率が高いペプチドである。

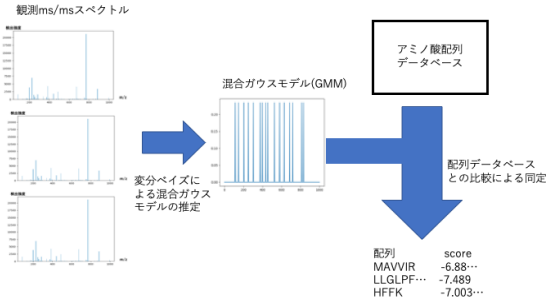


図 3　VBGMM によるペプチド配列推定

変分ベイズによる事後分布の推定としては、対数周辺尤度を最大化する事後分布を求めることでパラメータの事後分布を求める。つまり潜在変数 Z としたとき対数周辺尤度 $p(X) = \int p(X,Z)dZ$ を最大化するパラメータの事後分布を VB-EM アルゴリズム[5]により求める。最尤推定における二段階繰り返しの最適化の EM アルゴリズムにより事後分布を求める手法と異なる点は、事後分布を因子分解可能と仮定し近似分布で真の事後分布に近い分布を求めることである。VB-EM のアルゴリズムは事後分布を利用して潜在変数の事後分布を改良する E-step,潜在変数の事後分布によりパラメータの事後分布を改良する M-step を繰り返すことによって最適化を行う。これにより求めた混合ガウスモデルをアミノ酸配列のデータベースと比較しペプチド配列の同定を行う。

3. 結果の評価

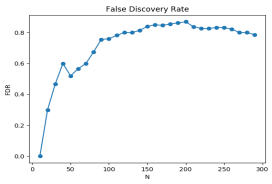
プロテオームは多数のタンパク質を対象とするため、解析の信頼性には多重検定が必要となる[6]。プロテオームなどの分野では FDR(False Discovery Rate)を求め、制御する手法が多く利用されている。FDR を利用する目的としては、ある程度の偽陽性(誤った同定された)を受け入れることで偽陰性、つまり同定できなかったタンパク質の数を減らすためである。しかし、FDR を算出するためには偽陽性の数が必要となる。したがって同定結果内の誤った配列に同定されてしまった結果の数を求められるようにする。その為の方法として、本来のタンパク質データベースに対して、本来の配列を逆にしたリバース配列、つまり実際には存在しないデコイ配列をデータベースに混ぜて同定を行う(Target-Decoy search[7])。これにより、本来ターゲットとする配列に同定されるはずが、誤ってデコイ配列に同定されてしまった数を求めることができる。これは本来デコイとする配列に同定されるはずが誤ってターゲットとするペプチドに同定されてしまう確率と等しくなる[6]。

FDR は次のように求められる。

FDR＝(デコイ配列の同定数 × 2)/(全ての同定数)

この FDR の値を計算することで同定の結果の偽陽性の占める割合がわかる。本研究における陽性は配列を同定できたことであり、陰性は同定できなかったことである。偽陽性は同定できたがその結果が間違っていることであり、誤った同定結果である。偽陰性は本来同定できるはずだったのに同定できなかった数である。

本手法によって生成した確率モデルを利用し 配列 DB からペプチドの推定を行った。その結果、計算したスコアが上位のペプチドと、タンパク質を同定する閾値を変化させた時の FDR の変化を図 4 に示す。グラフ(a)の横軸の N は同定したペプチドの数であり、縦軸は FDR であり(b)は本手法を用いてペプチドのスコアの計算を行なったものの一部を示している、



Sequence	Score	isdecoy
WPGR	−6.7111324	0
WVGR	−6.7111417	0
WGPR	−6.711591	0
WTGR	−6.711754	0
WGVR	−6.711950	0
WGTR	−6.712547	0
WSAR	−6.712713	0
WASR	−6.713327	0

(a)FDR　　(b)ペプチド同定スコア

図 4 提案手法による同定結果

4. まとめ、今後の課題

本研究では、プロテオーム解析における MS/MS のマススペクトルを無作為に抽出し VBGMM を用いた混合ガウス分布の推定を行うことで、マススペクトルを解析し、ペプチド配列の同定を行なった。また、同定結果を評価するための指標として、検索する配列にデコイ配列を加えることによって同定結果の FDR を計算し、同定結果にどれほどの誤りがあるかを求めた。今後の課題としては、実行時間の改善や、実際試料に含まれるタンパク質のペプチド配列が本手法同定でどの程度推測できているかを検討する必要がある。

謝辞

研究遂行にあたり、貴重な時間を割いて様々なアドバイスをいただきました国立循環器病研究センター研究所の病態ゲノム医学部 高橋篤 部長 及び 特任研究員 錦織充広 博士に感謝いたします。

参考文献

[1] Aebersold, Ruedi and Mann, Matthias, “Mass-spectrometric exploration of proteome structure and function”, Nature, Vol.537, 347.EP (2016).
[2] David N. Perkins,Darryl J.C. Pappin,David M.Creasy,John S.cottrell, “Probability‐based protein identification by searching sequence databases using mass spectrometry data”, Electrophoresis, 20, (18) (1999).
[3] Amos Bairoch, Brigitte Boeckmann, The SWISS-PROT protein sequence data bank, Nucleic Acids Research,vol19,2247-2249,(1991)
[4] Hagai Attias, “A Variational Bayesian Framework for Graphial Models”, Advances in neural information processing systems(2000)
[5] Bernardo, J. M., et al. "The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures." Bayesian statistics 7 (2003): 453-464.
[6] 吉澤 明康,どのデータベースを使うか ~データベース検索と配列解析・誤解と難題~,”Proteome Letters” 1,63-60,(2016)
[7] Elias, Joshua E and Gygi, Steven P, Target-decoy search strategy for increased confidence isn large-scale protein identifications by mass spectrometry, Nature Methods,4,207EP,02,(2007)