行政院國家科學委員會專題研究計畫 成果報告

蛋白質質譜資料質量校正之模擬研究
研究成果報告(精簡版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 95-2118-M-004-004-
執 行 期 間 ：95 年 08 月 01 日至 96 年 07 月 31 日
執 行 單 位 ：國立政治大學統計學系

計 畫 主 持 人：薛慧敏

計畫參與人員：博士班研究生-兼任助理：簡至毅、曾奕翔

處 理 方 式 ：本計畫涉及專利或其他智慧財產權，1 年後可公開查詢

中 華 民 國 96 年 12 月 07 日

行政院國家科學委員會補助專題研究計畫 ■ 成 果 報 告 □期中進度報告

蛋白質質譜資料之峰偵測與質量校正

計畫類別：■ 個別型計畫　　□ 整合型計畫
計畫編號：NSC 95－2118－M－004－004
執行期間：95 年 8 月 1 日至 96 年 7 月 31 日

計畫主持人：薛慧敏
共同主持人：
計畫參與人員：

成果報告類型(依經費核定清單規定繳交)：■精簡報告　□完整報告

本成果報告包括以下應繳交之附件：
□赴國外出差或研習心得報告一份
□赴大陸地區出差或研習心得報告一份
□出席國際學術會議心得報告及發表之論文各一份
□國際合作研究計畫國外研究報告書一份

處理方式：除產學合作研究計畫、提升產業技術及人才培育研究計畫、列管計畫及下列情形者外，得立即公開查詢
□涉及專利或其他智慧財產權，□一年□二年後可公開查詢

執行單位：

中　華　民　國　96 年　12 月　7 日

中文摘要

本計畫將針對蛋白質質譜圖(mass spectrum)資料，其事前處理(preprocessing)程序中的峰偵測(peak detection)與校正(alignment)步驟進行研究。此種資料主要來自 MALDI(Matrix Assisted Laser Desorption and Ionization)與 SELDI(Surface Enhanced Laser Desorption and Ionization)實驗。由此實驗得到生物樣本之質譜圖(mass spectrum, MS) ，圖中的橫軸為蛋白質之質量電荷比(mass-to-charge, m/z)，縱軸為其強度(intensity)之資料。由於實驗程序複雜，資料中充滿雜訊。此些雜訊將使研究人員無法正確偵測出蛋白質的出現以及適當的測量該蛋白質的強度。將導致在後續的資料分析，得到錯誤結論，故『峰偵測』與『峰校正』為兩個重要的事前處理(preprocess)程序。本研究將提出一新的處理方法，並將其應用於一組實際的 SELDI-TOF 資料上。我們發現此新方法相較於目前研究人員的常用方法—Bioconductor 的 PROcess 程式—要來的簡單並且有較佳的結果。

英文摘要

This research aims to study the alignment step in preprocessing the mass spectrometry (MS) data. Two popular mass spectrometry experiments are SELDI (Surface Enhanced Laser Desorption and Ionization) and MALDI (Matrix Assisted Laser Desorption and Ionization). In a mass spectrum of a biological sample, intensities of proteins are recorded in the vertical axis and the corresponding mass-to-charge (m/z) values are in the horizontal axis of proteins. Due to the complex nature of the experiment, the data is full of measurement error. The measurement error make it difficult to correctly identify the presence of a true feature and adequately quantify the strength of a present true feature. Peaks identification and alignment are thus two essential steps in data preprocessing. In this research, a new method for peak identification and alignment is proposed. A real SELDI-TOF mass spectra data set with known true features is used to assess the proposed method. This method is shown easy to implement and outperform the existing method in the Bioconductor PROcess package.

關鍵詞(keywords): Mass spectrometry, peak detection, peak alignment, continuous wavelet transformation.

前言

In identification of important biomarkers via proteomic mass spectra, procedures usually consist of two steps, namely, preprocessing step and analysis step. The preprocessing step usually includes baseline subtraction, normalization, peak detection, and alignment. In order to obtain good subsequent analysis results, an effective preprocessing step is required. Recent developments in analyzing microarray gene expression data have advanced the analysis step. However, there are still many issues in terms of the preprocessing step.

研究目的與文獻探討

Two critical steps for feature extraction of mass spectra are peak detection and peak alignment. Different procedures have been introduced and applied on MS data sets. Many studies suggest first to find peaks on individual spectrum and then to match and to align all the detected peaks subsequently (Tibshirani et al.(2004), Coombes et al.(2005), Morris et al.(2005), Yasui et al.(2003), Adam et al.(2002)). On the other hand, Morris et al.(2005), Wang, Cagney and Cartwright(2005)) considered to first form a reference spectrum, which integrates information of multiple spectra, and then to align and to detect peaks accordingly. Morris et al.(2005) compared the two aforementioned strategies through examples and simulation studies. They concluded that second strategy performs better overall. The noise in the reference spectrum is greatly reduced and the reference spectrum is more sensitive in finding peaks. From a virtual experiment, the reference spectrum algorithm has better performance in the sense of having greater sensitivity and less FDR.

The determination of occurrence of a peak is usually based on background-subtracted, normalized intensity value. A peak is detected if its intensity is relatively notable when compared with its neighborhood. Recently, Du, Kibbe, and Lin (2006) adopted a Continuous Wavelet Transform (CWT)-based procedure which utilizes not only the information from intensity but also the information from the shape of peaks through pattern matching in the wavelet space. Du et al.(2006) showed that this CWT-based algorithm can reduce false positive rate in detecting peaks. Even though the CWT-based peak detection algorithm has advantages over other algorithms, the CWT-based algorithm only deals with detecting peaks on one single spectrum at a time.

This paper aims to extend the CWT-based algorithm for peaks extraction on multiple spectra. We consider first pooling the shape information of all spectra in a reference and then identify the peaks based on the reference. Moreover, before integrating information, the experimental error in location is taken into consideration and reduced. The proposed method not only can perform peak detection for multiple spectra but also carry out peak alignment at the same time. The proposed method is called multi-spectra CWT-based(MCWT) algorithm.

研究方法

Assume there are n spectra. For the signal of spectrum i, $s_i(t)$, one first calculates the CWT coefficients,

$$C_i(a,b) = \int_R s_i(t)\varphi_{a,b}(t)dt, \quad \text{where } \varphi_{a,b}(t) = \frac{1}{\sqrt{a}}\varphi\left(\frac{t-b}{a}\right),$$

at scale a and translation b for i=1,…,n. Here $\phi(t)$ is the mother wavelet function. A CWT coefficients matrix is formed by taking consideration of a range of translation and various scales.

To reduce the deviation of location shift of the same peak, after the CWT coefficients are obtained, we take the local maximum over the window $b \pm \delta * b$,

$$C_i{}^*(a,b) = \max_{b' \in [b-\delta b, b+\delta b]} C_i(a,b'),$$

for each translation b, on spectrum i, at scale a. Since it is believed that the accuracy in the m/z position is within 0.3% of the m/z value, here δ=0.3%.

The next step is to integrate all the information on CWT coefficients matrices from n spectra into a single reference CWT coefficients matrix. For each a, b, the average values of the CWT coefficients over all samples, given by

$$C_{mean}{}^*(a,b) = \frac{\sum_{i=1}^{n} C_i{}^*(a,b)}{n}$$

are recorded in the reference CWT coefficients matrix. Linking the local maximal points at adjacent scales on the reference CWT coefficient matrix, one is able to find the ridge lines, which correspond to peaks with high possibility. Henceforth, peaks are identified based on the ridge lines found on the reference CWT coefficients matrix according to the criteria suggested by Du et al.(2006). The location of a peak is determined as the translation corresponding to the maximal CWT coefficient. For each individul spectrum, the identified peaks are then quantified as follows.

Assume there are p identified peaks with m/z values $b_{01},\ldots,b_{0p}$. The signal of the j-th identified peak for spectrum i is quantified as $S_{ij}=\max_a C_i^*(a,b_{0j})$.

結果與討論

In the following section, we consider a real data set with known polypeptide m/z positions for comparing our proposed method with another popular method, the PROcess package in Bioconductor. We consider the STANDARD data set provided by the organizers of the sixth international conference for the critical assessment of microarray data analysis (CAMDA, 2006, http://camda.duke.edu). The MS spectra was measured by Ciphergen NP20 chips. Twenty-one real peaks that are resulted by seven polypeptides with mass 7034, 12230, 16951, 29023, 46671, 66433 and 147300 with up to three charges are in the sample, see Du et al.(2006). The 32 spectra of high laser energy are used on assessment for an alignment procedure. Moreover, the spectra data with m/z values < 2k are ignored in subsequent analysis due to the fact that spectra data at low range m/z are too noisy for the machine to record stably.

Our MCWT-based algorithm and PROcess have different definitions for SNR. Thus different ranges of thresholds of SNR are employed for the two methods to obtain the ROC curves. Our methods take the SNR threshold ranged from 0.5 to 11 with increment of 0.3, while the PROcess takes the SNR threshold from 1 to 11 with increment of 0.5. In PROcess, default setting is used for background subtraction and normalization before peak detection and alignment. On the contrary, under mild assumptions, the two preprocessing steps are unnecessary for the MCWT-based algorithm method.

Our methods dominate the method of PROcess package in terms of having lower FDR and higher sensitivity. Also one finds that the number of identified peaks, sensitivity and FDR of the MCWT-based algorithm are monotonic decreasing as the SNR threshold. Less intuitively, such monotonicity does not exist in PROcess.

REFERENCES

- ✧ Adam, B.-L., Qu, Y., Davis, J. W., Ward, M. D., Clements, M. A., Cazares, L. H., Semmes, O. J., Schellhammer, P. F., Yasui, Y., Feng, Z. and Wright, G. L., Jr. (2002) Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. Cancer Research, 62, 3609-3614.

- CAMDA (2006) CAMDA 2006 Competition Data Set.
- Coombes, K. R., Tsavachidis, S., Morris, J. S., Baggerly, K. A., Hung, M.-C. and Kuerer, H. M. (2005) Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. Proteomics, 5,4107-4117.
- Du, P., Kibbe, W. A. and Lin, S. M.(2006) Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. Bioinformatics, 22, 2059-2065.
- Gentleman, R. and Vandal, A. C.(2001) Computational algorithms for censored data problems using intersection graphs. Journal of Computational and Graphical Statistics,10, 403-421.
- Morris, J. S., Coombes, K. R., Koomen, J., Baggerly, K. A. and Kobayashi, R. (2005) Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. Bioinformatics, 21, 1764-1775.
- Tibshirani, R., Hastie, T., Narasimhan, B., Soltys, S., Shi,G., Koong, A. and Le, Q.-T. (2004) Sample classification from protein mass spectrometry, by 'peak probability contrasts'. Bioinformatics, 20, 3034-3044.
- Wong, J. W. H., Cagney, G. and Cartwright, H. M. (2005) SpecAlign-processing and alignment of mass spectra datasets. Bioinformatics Applications Notes, 21, 2088-2090.
- Yasui, Y., Pepe, M., Thompson, M. L., Adam, B.-L., Wright, G. L., Jr., Qu, Y., Potter, J. D., Winget, M., Thornquist, M. and Feng, Z. A data-analytic strategy for protein biomarker discovery:profiling of high-dimensional proteomic data for cancer detection.Biostatistics, 4, 449-463.