A novel approach for analysis of LC/MS data - Peak Clustering and Fitting

LC/MS 데이터 분석의 새로운 접근 방법 - 피크 군집화와 조정

Byung Hwa Lee* and Joon Hee Han

Department of Computer Science and Engineering, POSTECH, Pohang, Republic of Korea

***To whom correspondence should be addressed. E-mail: lbh76@postech.ac.kr**

Abstract

LC/MS를 이용하여 펩타이드 혹은 단백질 같은 물질을 분석하는 실험이 급격히 늘어남에 따라 LC/MS 데이터를 자동으로 처리하는 기술에 대한 요구가 커지고 있다. 이러한 LC/MS 데이터의 자동 분석 기술에 대한 연구는 현재 활발히 진행되어 왔고, 이를 직접 구현한 여러 상용 소프트웨어들이 개발되어 있는 상태이다. LC/MS 데이터는 noise 제거, background 데이터 제거, deconvolution 알고리즘을 적용한 분자량(molecular weight) 할당 등의 작업을 거쳐 분석하게 된다. 이러한 과정을 거쳐 얻어진 분자량에 대한 데이터가 올바른 값인지 검증하는 작업이 필요하다. 본 논문에서는 이러한 검증 작업과 관련하여 Peak Clustering and Fitting(이하 PC&F)에 대한 알고리즘을 제안한다. PC&F은 peak 데이터들이 지니고 있는 속성에 대한 Mahalanobis distance를 이용하여 peak 데이터를 각 retention time에 따라 clustering 분석을 하는 작업이다. 본 논문에서 제안하는 PC&F 알고리즘을 Microsoft Visual C++ 6.0 MFC 환경에서 직접 개발한 소프트웨어(PeakClusterFitLCMS)로 실험하였다. 실험결과 PC&F 작업을 통해 동일한 구성물질로부터 발생한 peak 데이터를 모아서 보다 신뢰할 수 있는 분자량을 구할 수 있었고, 구성물질에 의해 발생되지 않은 noise peak 데이터를 찾아 제거시킬 수 있음을 확인할 수 있었다.

Introduction

Liquid chromatography /mass spectrometry(LC/MS)는 단백질 혹은 펩타이드와 같이 매우 작은 물질을 대용량 처리(high-throughput)로 분석하는 기술로

본 연구는 과기부의 특정연구 개발사업 (National R&D Program-Fusion Strategy of Advanced Technologies) 으로 부터 지원 받았음.

써 최근에 널리 이용되고 있다[1]. 이러한 LC/MS 분석은 하나의 생물학적인 물질 내에 존재하는 구성물질들의 질량 분석을 통해 이루어진다. LC/MS를 이용하여 분석하고자 하는 물질을 제대로 인식하기 위해서는 LC/MS에 의해 얻어지는 데이터를 정확하게 분석하는 기술이 선행되어야 한다.

최근에 컴퓨터를 이용하여 LC/MS 데이터를 보다 효율적이고, 정확하게 분석하는 기술이 많이 연구되고 있다. 이러한 기술은 LC/MS 데이터 내부의 noise 제거, background 데이터 제거, 데이터 내의 peak 발견, deconvolution 알고리즘을 이용한 구성물질들의 분자량 할당 등의 작업으로 나누어질 수 있다[3]. 이러한 일련의 과정을 거쳐 얻어진 분자량이 여러 요인에 의해 잘못된 결과를 낳을 수 있다. 그 예로 각 retention time에 따라 분자량이 되는 peak의 위치가 조금씩 어긋나는 현상과 실제 구성물질의 분자량이 아닌 곳에서 peak 위치가 나타나는 현상 등을 들 수 있다. 이러한 잘못된 결과를 바로잡기 위해 최종적인 검증 작업이 필요하다. 본 논문에서는 위와 같은 일련의 LC/MS 데이터 분석과정 중 결과 데이터에 대한 검증과 관련하여 새로운 접근 방식인 PC&F 알고리즘에 대하여 기술한다.

PC&F 작업은 구성물질의 분자량이 되는 peak 데이터들을 retention time에 따라 clustering 분석을 하는 작업이다. 이러한 작업을 통해 하나의 구성물질로부터 발생한 peak 데이터를 모아서 보다 신뢰할 수 있는 대표값을 얻을 수 있고, 에러에 의해 잘못 구해진 peak 데이터를 찾아 제거시킬 수 있다. 이러한 PC&F 작업은 패턴인식의 Mahalanobis distance를 이용한 clustering 기술을 이용하여 구현했다.

Methods

기존의 LC/MS 데이터 분석시스템 및 연구

LC/MS 분석은 상당히 많은 양의 데이터를 생성하게 된다. 하나의 혼합물질을 분석하는 데 생성되는 데이터가 10,000 mass spectra이상이 되는 것이 보통이다. 이렇게 많은 양의 데이터를 보다 효율적으로 분석하기 위해서는 컴퓨터를 이용한 소프트웨어 분석 시스템 개발이 필수적이다[3]. 현재 많은 LC/MS 데이터 분석 소프트웨어가 개발되어 있는 상태이고, 이러한 소프트웨어는 대개 상용화된 LC/MS 장비의 일부분으로 제공되고 있다[4]. LC/MS 데이터 분석 소프트웨어 중 하나인 ProMAPDB는 data collection, peak identification, spectral purification, mass spectral integration of scans in a peak, assignment of molecular weights for observed proteins by using a deconvolution algorithm 등과 같은 과정을 거쳐 데이터를 분석하게 된다. 이 소프트웨어에서는 분석하고자 하는 물질의 데이터와 기존의 데이터가 저장되어있는 database를 비교 검색하는 기능을 제공함으로 써 구성물질을 보다 효율적으로 인식하는데 도움을 주고 있다[3].

LC/MS 데이터 분석 기법과 관련된 연구 또한 활발히 진행되어 왔다. 대표적으로 LC/MS 데이터의 noise를 없애기 위한 CODA 알고리즘[6], 하나의 물질로부터 나온 모든 peak 데이터들을 하나의 singly charged monoisotopic peak으로 만들기 위한

deconvolution and deisotoping 알고리즘[2]에 대한 연구 등을 들 수 있다. LC/MS 데이터의 자동분석기술에 대한 연구는 그 수요의 증가에 따라 보다 활발해지고 있다.

Peak Clustering and Fitting (PC&F)

Peak Clustering and Fitting (PC&F)이란 일련의 과정을 거쳐 얻어진 peak 데이터에 대해 retention time 축을 따라 clustering을 수행하고 각 cluster에 포함된 peak 데이터들을 보다 신뢰할 수 있는 대표값으로 맞추어 주는 작업으로 정의 될 수 있다. 이 작업의 주 목적은 한 구성물질로부터 발생한 peak 데이터들을 하나의 cluster로 모으고, 모아진 데이터를 바탕으로 보다 정확한 분자량을 얻는 데 있다. 또한 cluster에 포함되지 않은 noise peak 데이터를 찾아 제거하기 위함이다. 이러한 작업은 LC/MS 데이터를 효율적으로 분석하는 데 있어 효과적인 방법으로 본 논문에서 실험으로 검증해 보았다.

PC&F 작업이 필요한 이유는 다음과 같다. 우선 peak 데이터에서의 mass값이 각 retention time 에서 실험적, 기기적 요인에 의해 조금씩 어긋나는 현상이 발생하기 때문이다. 이는 각 retention time에서 mass spectrum을 따로 따로 뽑아내는 LC/MS 분석기기의 방식에 의해 발생한다. 이러한 현상은 PC&F 작업을 통해 peak 데이터의 cluster를 구하고 그 cluster 내의 모든 데이터의 mass값에 대한 평균(대표값)을 그 cluster내의 전체 데이터 값으로 맞추어 줌으로써 바로 잡을 수 있다.

또 다른 이유는 noise 데이터에 의해서 실제 구성물질의 분자량이 아닌 값에서 peak 데이터가 형성될 수 있기 때문이다. 이러한 peak들은 retention time축에서 연속되게 나타나지 않고 특정 retention time 상에서만 홀로 발생하게 된다. 따라서 PC&F 작업을 통해 얻은 결과에서 어떠한 cluster에도 포함되지 않은 peak 데이터들을 제거시킴으로써 바로 잡을 수 있다.

특히 후자에 대한 효과는 본 논문의 실험을 통해 확실히 확인할 수 있었다. 즉 PC&F 작업을 거치지 않은 데이터의 경우 실제 구성물질이 아닌 noise peak 데이터에 의한 바르지 못한 분자량까지도 구성물질의 분자량으로 잘못 인식할 수 있는 반면 PC&F작업을 적용한 경우 구성물질로써 확실히 검증된 분자량 만을 뽑아낼 수 있다. 이는 정확한 LC/MS 데이터 분석에 큰 도움을 준다.

PC&F작업은 여러 분석 작업을 거친 결과 데이터 상에서 이루어지는 peak 데이터를 검증하는 부분이다. 이러한 작업은 다른 논문들에서는 자세하게 언급되어 있지 않지만 참고논문[4]에서 간단히 제시되어 있다. 이 논문에서는 마지막으로 얻어진 mass peak 값들 상에서 간단한 소팅(sorting) 작업을 통해 수행하고 있다[4]. 하지만 본 논문에서는 보다 많은 요소를 고려하여 구현하였다.

PC&F의 적용 시기

전체 LC/MS 데이터 분석 과정 중 PC&F 알고리즘을 적용할 수 있는 시기는 두 가지이다. 첫 번째는 noise제거, background 데이터 제거 과정을 끝낸 후 m/z상에서 PC&F 작업을 수행하는 것이다. 이 경우에는 PC&F을 통해 검증된 m/z 축 상의 데이터를 얻은 후 deconvolution 알고리즘을 적용하여 최종적

인 구성물질에 대한 분자량을 얻게 된다. 두 번째는 noise제거, background 데이터 제거, deconvolution 알고리즘까지 적용한 후 그 결과 데이터상에서 PC&F을 마지막 검증작업으로 수행하는 것이다.

본 논문에서는 후자의 상황 즉 deconvolution까지 끝마친 상태의 데이터를 이용하여 PC&F알고리즘을 적용하는 실험을 수행하였다. 여기서 deconvolution된 데이터는 각 retention time 의 mass spectrum에서 개별적으로 deconvolution한 결과를 의미한다. 이 데이터를 이용하여 여러 retention time을 따라 PC&F 알고리즘을 적용시켜 보았다. 향후 전자의 상황에서 적용시킨 효과에 대해서도 실험해 볼 계획이다.

Peak 데이터의 Clustering

패턴인식에서 clustering이란 ' 유사한 속성을 지니고 있는 데이터를 cluster 혹은 group이라는 특성으로 묘사하는 작업' 으로 간단히 정의하고 있다[5]. 본 논문의 주제인 PC&F 또한 peak 데이터의 속성에 따라 군집화하는 작업이므로 이러한 clustering 기술을 직접적으로 이용해야 한다. 데이터를 clustering하는 기준으로 다음과 같은 두 가지를 들 수 있다[5].

- 샘플데이터들 간의 유사성 측정을 기준으로 한 clustering
- cluster들 내의 샘플 데이터 집합의 분리 시 상태 평가에 의한 clustering

PC&F 작업에서는 위 두 가지 기준 중 첫 번째 기준을 적용하는 방식이 더욱 적절하다고 판단된다.

샘플 데이터 간의 distance는 유사성 측정의 좋은 기준이 되기 때문에 clustering기술에 많이 이용되고 있다. 여기서 사용되는 distance는 Euclidean distance, Minkowski metric, City block distance, Mahalanobis distance등이 있다[5]. 이러한 distance들 중 이 문제에 가장 적합한 distance를 선택하기 위해서는 clustering을 수행하는 대상인 LC/MS peak 데이터의 특성을 잘 살펴볼 필요성이 있다.

본 논문에서 다루고 있는 LC/MS peak 데이터는 mass와 area(혹은 intensity) 두 가지 feature를 가지고 있는 2차원 데이터이다. 또한 이 데이터의 area feature는 mass feature에 비해 분산이 훨씬 크게 분포되어 있다. 이러한 데이터 특성으로 인해 Euclidean distance, City block distance 보다는 데이터들 간의 feature covariance 특성이 반영된 Mahalanobis distance를 이용하여 유사성 정도를 측정하는 것이 가장 적절하다. 다음 식은 두 peak데이터 간의 유사도를 측정하기 위한 Mahalanobis distance의 제곱을 나타내는 식이다.

$$r^2 = (x_1 - x_2)^t \sum^{-1} (x_1 - x_2)$$

$\sum$: mass를 기준으로 한 threshold내에 있는 peak sample들의 covariance matrix

x_1 : 비교의 기준이 되는 peak sample data

x_2 : 비교하고자 하는 peak sample data

본 논문에서 제안하는 PC&F에서는 peak 데이터의 mass 값을 기준으로 일정 threshold 내부에 있는 peak 데이터만을 고려하여 clustering한다. 이는 특정 mass값을 넘어서는 peak 데이터들은 동일한 구성물질로부터 발생한 데이터가 될 수 없는 LC/MS데이

터 속성을 이용한 것으로 모든 peak 데이터 상에서 clustering을 수행하는 것에 비해 훨씬 효율적이다. PC&F에 대한 자세한 알고리즘은 다음 장에 자세히 소개되어 있다.

PC&F 알고리즘

본 논문에서 제안하는 알고리즘은 하나의 구성물질에 대한 peak cluster를 찾는 부분과 전체 구성물질에 대한 peak cluster를 찾는 부분으로 나눌 수 있다. 여기에서 후자는 전자의 알고리즘을 이용하여 구현할 수 있다.

하나의 구성물질에서 PC&F 수행 알고리즘

A. PC&F 하고자 하는 한 구성물질의 peak 데이터가 주어지게 되면 그 peak 데이터를 초기 peak 데이터 값으로 놓는다.

B. 초기 peak 데이터의 mass 값을 중심으로 threshold 값을 적용하게 된다. 앞으로는 이 threshold값 범위내의 peak 데이터들 만 고려하여 clustering을 수행한다.

C. Threshold 내에 존재하는 모든 peak 데이터 샘플들의 covariance matrix를 구한다.

D. 초기 peak 데이터로부터 retention time이 작아지는 방향으로 가장 유사하다고 판단되는 peak 데이터를 하나씩 뽑아낸다. 이때 유사성을 측정하는 기준은 peak 데이터들간의 Mahalanobis distance이고, 유사성을 측정하는 대상은 바로 전 단계에서 clustering된 peak 데이터와 현재의 retention time상에서 threshold 내의 모든 peak 데이터들 간이다. Mahalanobis distance가 가장 작은 peak 데이터를 뽑아내어 cluster내에 포함시킨다. 여기서 Mahalanobis distance의 covariance는 'C'에서 구해진 값이다. Mahalanobis distance가 가장 작은 데이터라도 사용자에 의해 주어지게 되는 일정 Mahalanobis distance를 초과하는 경우 그 데이터는 cluster에 포함시키지 않는다.

E. 'D'와 같은 과정으로 clustering을 수행하는 도중 사용자에 의해 지정되는 조건을 갖추지 못하는 경우 clustering을 종료하게 된다. 여기서 사용자에 의해 지정될 수 있는 조건은 cluster 내 peak 데이터에서 retention time 상으로 연속되게 나타나지 않는 retention time scan 수 이다. 즉 clustering되는 peak 데이터가 일정 retention time scan수 만큼 연속적으로 나타나지 않을 경우 clustering을 종료시킨다.

F. 'D', 'E'의 단계를 초기 peak 데이터로부터 retention time이 커지는 방향으로 반복 실행한다.

G. 얻어진 peak cluster 중 사용자에 의해 지정되는 조건을 갖추지 못하는 peak cluster는 하나의 구성물질로 인정되지 않는다. 여기서 사용자에 의해 지정될 수 있는 조건은 cluster내의 총 peak 데이터 개수이다. 즉 cluster 내에서 peak 데이터수가 일정 값을 초과하지 않는 경우 그 cluster 값은 분자량에서 제외시킨다. 이는 retention time 축을 따라 연속적으로 그 peak 데이터가 나와야지만 하나의 구성물질로 판단될 수 있기 때문이다.

H. clustering된 peak 데이터 상에서 mass에 대한 대표값을 구하여 그 값을 최종적인 분자량으로 할당하게 된다. 여기서 생각할 수 있는 대표값은 크게 네가지를 들 수 있다. clustering된 mass값들의 평균값, area값에 대해 weight를 주는 평균값, 많이 발

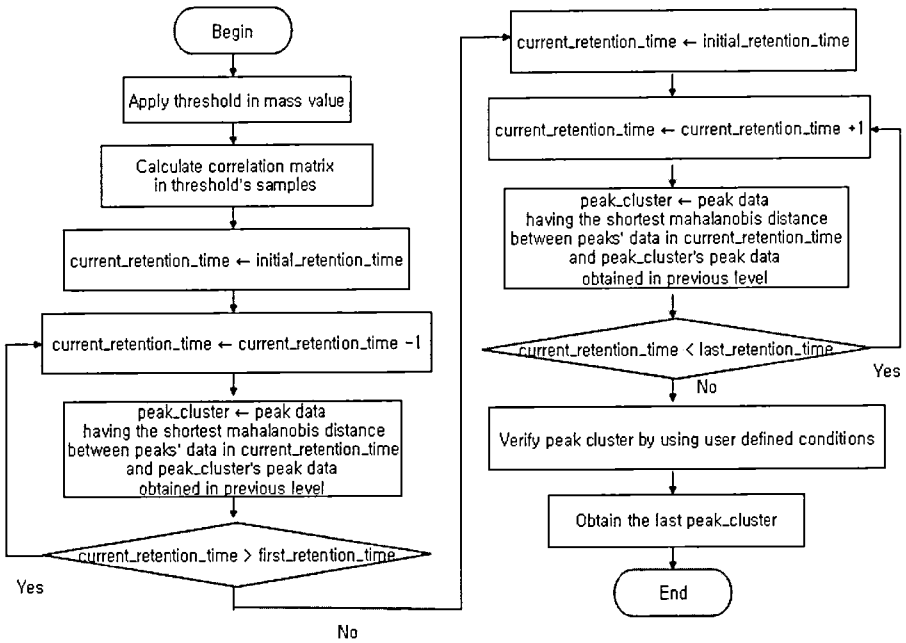

그림 1. 하나의 구성물질에서 PC&F 수행 알고리즘에 대한 순서도

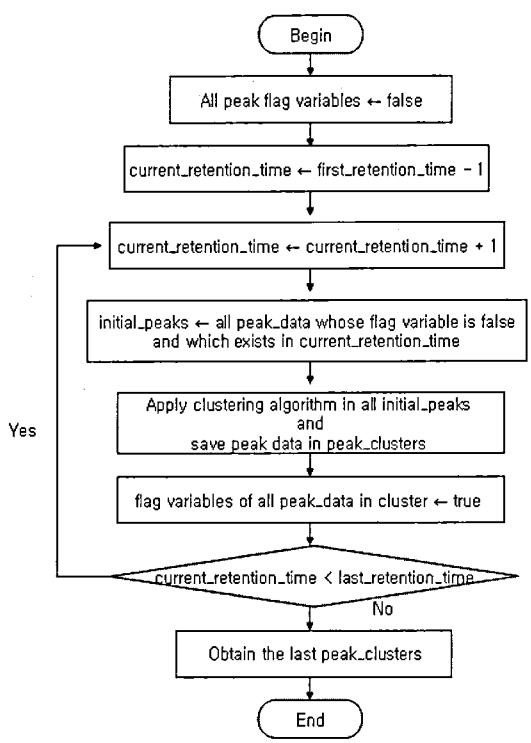

그림 2. 전체 LC/MS 데이터에서의 PC&F 수행 알고리즘에 대한 순서도

견되는 빈도에 따른 최빈값, 그리고 마지막으로 mass 값에 대한 중간값이다. 본 논문에서는 이 네 가지 값을 적용시켜 실험해 보았다.

지금까지 설명한 PC&F 알고리즘에 대한 대략적인 순서도를 그림 1에서 참조할 수 있다.

전체 LC/MS 데이터에서의 PC&F 수행 알고리즘

A. 분석하고자 하는 LC/MS 데이터에서 가장 작은 retention time 축 상에 있는 peak 데이터들을 초기 peak 데이터로 지정한다.

B. 각 초기 peak 데이터에서 앞에서 설명한 '하나의 구성물질에서 PC&F 수행 알고리즘'을 적용하여 clustering을 수행한다. 이렇게 clustering을 수행하면서 하나의 구성물질로 clustering된 peak 데이터는 그 데이터가 cluster에 포함되어 있는 peak 데이터임을 인식할 수 있도록 flag변수를 이용하여 기록하게 된다. 이러한 peak 데이터들은 후에 다른 peak cluster에 포함될 수 없게 하고, 다음 단계에서 초기 peak 데이터로 지정될 수 없도록 해야 한다.

C. 위와 같은 방식으로 가장 작은 retention time축에서 clustering이 끝나면 두 번째 retention-time 축에서 cluster에 포함되지 않은 peak 데이터들을 flag 변수를 통해 확인하여 초기 peak 데이터로 지정하고 'A'와 'B' 작업을 반복하게 된다. 이와 같은 과정을 가장 큰 retention time 축 까지 반복하여 수행하여 알고리즘을 마친다.

위에서 설명한 알고리즘을 나타내는 순서도를 그림 2에서 참조할 수 있다.

Results

PeakClusterFitLCMS 소프트웨어

지금까지 논의한 PC&F 알고리즘을 Microsoft Visual C++ 6.0 MFC 환경에서 소프트웨어로 구현하였다. 이 소프트웨어는 일련의 LC/MS 데이터 분석과정에서 deconvolution까지 마친 데이터를 입력으로 받아 데이터를 분석한다. 입력 데이터는 mass, area, retention time feature를 가지는 3차원 peak 데이터이다. 이 데이터는 소프트웨어 상에서 graph로 출력함으로써 시각적으로 분석하기 용이하도록 구현하였다. 출력되는 graph에서 x축은 mass, y축은 retention time을 나타내고 graph상의 점은 각 peak의 데이터 값을 나타낸다. peak 데이터들의 정확한 수치는 대화상자를 통해 출력함으로 써 손쉽게 확인할 수 있는 기능을 제공한다. 그리고 graph에서 확대, 축소 기능을 제공하여 많은 양의 데이터가 입력될 때 분석이 용이하도록 구현했다.

PC&F은 앞서 설명한 알고리즘을 그대로 구현했다. 여기서 PC&F은 두 가지 형태의 기능을 제공한다. 첫 번째는 하나의 peak 데이터에 대해서 수행하는 것이고, 두 번째는 입력 받은 전체 LC/MS 데이터에서 수행하는 것이다. PC&F의 결과는 대화상자를 통해 수치 상으로 출력하도록 했고, 시각적으로 분석이 용이하도록 graph상에서도 표현되도록 했다. PC&F알고리즘을 적용하여 얻은 하나의 cluster는 모아진 peak 데이터를 선으로 연결시켜 표현했다. 또한 PC&F을 통해 얻어진 대표값(검증된 구성물질의 분자량)으로 조정된 peak 데이터에 대한 graph를 출력하

PeakClusterFitLCMS 소프트웨어 실행화면(그림 3, 그림 4)

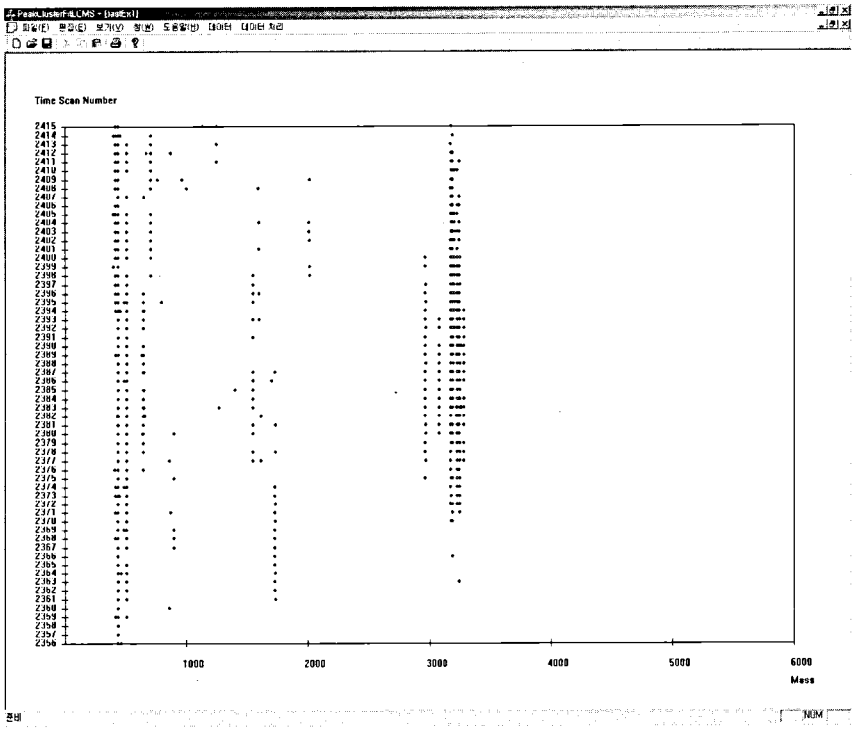

그림 3. peak 데이터를 입력 받아 graph상에 출력한 화면

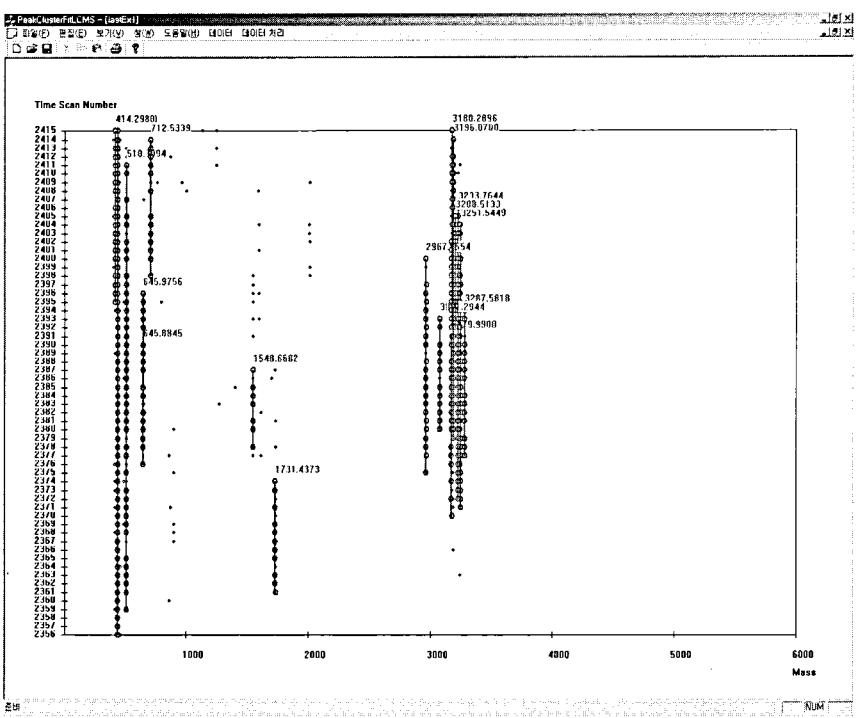

그림 4. PC&F실행결과

는 기능도 제공한다.

앞에서 설명한 PC&F 알고리즘에서 사용자가 지정해야 할 여러 초기값들은 대화상자를 통해 쉽게 입력할 수 있도록 구현했다.

실험결과

구현한 PeakClusterFitLCMS 소프트웨어를 이용하여 PC&F 작업의 유용성을 실험해 보았다. 실험에 사용된 LC/MS 기기는 'Applied Biosystem' 사의 'Q-STAR' 를 이용하였다. 이 기기는 Nano LC 와 Q-TOF 방식으로 LC/MS 실험을 수행하는 장비이다. des-Arg1-Bradykinin, Angiotensin I, Glu1-Fibrinopeptide B, ACTH (clip 1-17), ACTH (clip 7-38) 이 포함되어 있는 혼합물질에 대해 'Q-STAR' 기기를 이용하여 LC/MS 분석 데이터를 얻었다. 이 물질들은 각 각 904.4681, 1296.6953, 1570.0867, 2093.0867, 3657.9294 의 분자량을 가지고 있는 물질이다(단위: amu(atomic mass unit)).

이 LC/MS 데이터를 대상으로 본 논문에서 제안한 PC&F 알고리즘을 적용시킨 결과와 적용시키지 않았을 때의 결과를 비교 분석해 보았다. PC&F 작업을 하기 전 일련의 분석작업은 'Q-STAR' 장비에서 기본적으로 제공되는 소프트웨어인 'Analyst QS' 라는 프로그램을 이용하였다. 그리고 PC&F 알고리즘을 적용시킬 때 사용자에 의해 지정되는 조건은 threshold 로써 clustering 하고자 하는 peak 데이터의 mass 값을 기준으로 -3, +3 범위 이내를, 연속되게 나타나지 않은 retention time scan 수로 3 이하를, 허용할 수 있는 Mahalanobis distance 를 9 이하를, 구성물질로 인정할 수 있는 cluster 내의 peak 데이터 개수를 5 이상으로 놓고 수행했다.

표 1 은 PC&F 작업을 통하여 위의 구성물질들에 대한 분자량을 구한 실험결과이다. 표에 나타나 있는 여러 수치는 PC&F 적용 시 한 cluster 에 대한 대표값으로 어떠한 값을 적용시키느냐에 따른 결과들이다. 여기에서 실제값은 이전에 알려져 있는 각 구성물질의 분자량이고, 평균값은 PC&F 작업을 통해 얻은 한 cluster 내의 mass 값들의 평균, 최빈값은 mass 값들의 빈도가 가장 많이 나타나는 범위의 평균, 중간값은 mass 값들의 크기 상에서 중간에 위치하는 값, area 값에 대한 가중치를 둔 평균값은 각 peak 데이터의 area feature 값에 가중치를 고려하여 계산한 mass 평균값을 의미한다. 실험결과에서 실제값과 결과값이 차이가 나는 이유는 실험 시 환경 혹은 기기상태에 따라서 오차가 발생한 것이다. 이 실험결과에서는 평균값을 사용하는 것이 실제값과 가장 가깝게 나타났지만 실험기기의 오차를 고려하지 않았으므로 어떠한 값이 가장 적절한지 확실히 단정하기 어렵다. 하지만 이 실험을 통해 PC&F 적용했을 때 실제 구성물질의 분자량과 유사한 값을 얻을 수 있었다. 이는 PC&F 알고리즘을 적용시켜 실제 구성물질의 분자량을 찾을 수 있음을 의미한다. 보다 많은 실험을 통해 어떠한 대표값이 가장 정확한지 확인할 계획이다.

표 2 는 PC&F 알고리즘을 적용시켰을 때와 적용시키지 않았을 때 고려해야 할 peak 데이터 개수와 그 개수가 전체 데이터의 개수에서 차지하는 비율이다. 표에 나와있

구성물질 명	실제값	평균값	최빈값	중간값	Area 값에 대한 가중치를 둔 평균값
des-Arg1-Bradykinin	904.4681	903.419377	903.415	903.4185	903.414441
Angiotensin I	1296.6953	1295.591103	1295.585	1295.5895	1295.590346
Glu1-Fibrinopeptide B	1570.0867	1569.514539	1569.525	1569.5177	1569.508825
ACTH (clip 1-17)	2093.0867	2092.267427	2091.955	2091.9621	2092.244597
ACTH (clip 7-38)	3657.9294	3657.620190	3657.755	3657.7011	3657.578512

표 1. PC&F 알고리즘을 이용하여 찾은 대표값에 따른 분자량 (단위: amu)

Retention time	11.218 min		22.469 min		32.319 min	
	고려해야 할 mass 개수	전체 데이터 내의 비율	고려해야 할 mass 개수	전체 데이터 내의 비율	고려해야 할 mass 개수	전체 데이터 내의 비율
PC&F 비 적용 시	619 개	100%	769 개	100%	498 개	100%
PC&F 적용 시	29 개	4.7%	33 개	4.3%	25 개	5.0%

표 2. 각 retention time 부근에서 PC&F 알고리즘을 적용시켰을 때와 적용시키지 않았을 때 분자량을 구하기 위해 고려해야 할 peak 데이터 개수와 그 개수가 전체 데이터의 개수에서 차지하는 비율

Retention time	11.218 min	22.469 min	32.319 min	43.437 min
전체 데이터에서 noise peak 비율	19.7%	21.1%	25.5%	26.4%

표 3. 각 retention time 부근에서 전체 peak 데이터 내에서 PC&F 작업으로 제거된 noise peak 데이터가 차지하는 비율

는 각 retention time 부근에서 30 retention time scan line 상에서 PC&F 알고리즘을 적용한 데이터를 이용하였다. PC&F 알고리즘을 적용하기 전에는 모든 peak 데이터의 mass 값을 고려하여 구성물질의 분자량을 찾아야 하지만 PC&F 알고리즘을 적용한 후에는 동일한 구성물질로부터 발생한 peak 데이터들을 하나의 cluster 로 모으고, 대표값으로 맞추어졌기 때문에 그 대표 mass 값만 고려하여 분자량을 찾으면 된다. 이 실험결과를 통하여 PC&F 알고리즘이 분자량을 구하기 위해 분석해야 할 데이터

양을 매우 효과적으로 줄여줄 수 있음을 확인할 수 있었다.

마지막으로 표 3 에서는 PC&F 작업을 수행했을 때 noise peak 데이터 제거에 대한 효과를 실험하였다. 앞서 살펴보았듯이 PC&F 알고리즘을 적용시키면 전체 peak 데이터에서 noise peak 데이터를 찾아 제거할 수 있다. 실험결과는 각 retention time 부근에서 noise 로 제거되는 peak 데이터의 비율을 나타낸다. 평균적으로 전체 데이터 상에서 23.2%의 noise peak 데이터를 제거시키는 효과를 얻을 수 있었다.

Discussion

지금까지 LC/MS 데이터를 분석하는 데 있어 본 논문에서 제안한 방식인 PC&F 알고리즘에 대해 살펴보았다. PC&F 알고리즘을 통해 LC/MS 데이터에서 한 물질에 대한 peak 데이터를 하나의 cluster로 모을 수 있었고, 그 cluster의 대표값을 구함으로 써 데이터를 보다 효율적이고 정확하게 분석 할 수 있었다. 또한 noise peak 데이터를 제거하는 데도 큰 효과를 얻을 수 있었다. 본 논문에서는 deconvolution 과정까지 끝낸 LC/MS 데이터를 이용해서 실험했지만, 앞으로는 deconvolution 과정 전에 PC&F 알고리즘을 적용한 후 deconvolution 을 수행하는 실험을 진행할 계획이다.

Acknowledgements

본 연구는 과기부의 특정연구 개발사업 (National R&D Program-Fusion Strategy of Advanced Technologies) 으로 부터 지원 받았습니다. 그리고 실험자료를 제공해주신 포항공대 생명과학과 김윤동 연구원에게 감사의 말씀을 전합니다.

References

[1] Curtis A. Hastings, New algorithms for processing and peak detection in liquid chromatography/mass spectrometry data, *Rapid Commun. Mass Spectrom.*, 2002, 16, 462-467

[2] Marco Wehofsky, Automated deconvolution and deisotoping of electrospray mass spectra, *J. Mass Spectrum.*, 2002, 37, 223-229

[3] Samir V. Deshpande, Automated and rapid bacterial identification using LC-Mass Spectrometry with a relational database management system, *2004 IEEE Computational Systems Bioinformatics Conference,* 472-473

[4] John O. Pearcy, MoWeD, a computer program to rapidly deconvolute low resolution electrospray liquid chromatography/mass spectrometry runs to determine component molecular weights, *J Am Soc Mass Spectrom.*, 2001, 12, 599-606

[5] Richard O. Duda, Pattern Classification, *Wiley-interscience,* 2001, Second Edition, 537-542

[6] Willem Windig, A noise and background reduction method for component detection in liquid chromatography/mass spectrometry, *Anal. Chem.*, 1996, 68, 3602-3606