

# A New Method for Detecting Isotope Patterns in Liquid Chromatography/Mass Spectrometry Data

## LC/MS 데이터에서 동위 원소 패턴을 찾는 새로운 방법

Ji-Woon Hwang<sup>1\*</sup>, Youndong Kim<sup>2</sup>, Joon Hee Han<sup>1</sup>

<sup>1</sup> Department of Computer Science and Engineering, POSTECH, Pohang, Korea

<sup>2</sup> Department of Life Science, POSTECH, Pohang, Korea

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

### Abstract

본 논문에서는 LC/MS 데이터로부터 동위 원소 패턴(isotope pattern)을 찾는 새로운 방법을 제시하고자 한다. 기존의 분석 방법에서는 LC/MS 데이터를 1차원적으로 분석하고 있기 때문에 2차원에서 적용할 수 있는 알고리즘을 적용하기가 어렵다. LC/MS 데이터를 2차원 영상으로 가시화해 본 결과, 하나의 동위 원소 패턴에 속하는 단일 동위 원소 피크(single isotope peak)는 모양, 크기와 같은 2차원 형태적 특징들도 유사함을 알 수 있다. 따라서, 기존의 방법들이 질량 스펙트럼과 같은 1차원 신호를 분석하는 것에 중점을 둔 것에 비해, 본 논문에서는 LC/MS 데이터를 2차원 신호 즉, 영상(image)으로 간주하고 영상 처리 방법과 객체 인식 방법을 적용하였다. 실험 결과 같은 동위 원소 패턴에 속하는 각각의 단일 동위 원소 피크들 사이에 peak maxima position 뿐만 아니라 skewness, variance등도 유사였으며 이러한 유사도를 기반으로 동위 원소 패턴을 찾을 수 있었다.

### Introduction

크로마토그래피(chromatography)와 질량 분석기(mass spectrometry)는 다양한 분야에서 유기화합물의 분석을 위해 널리 사용되어 왔다. 근래에 액체 크로마토그래피(liquid chromatography)와 질량 분석기의 조합은 더욱 강력한 분석 능력을 가져다 주었으며, 이러한 LC/MS 분석 방법은 고분자 화합물의

분자량 측정에 널리 이용되고 있다.

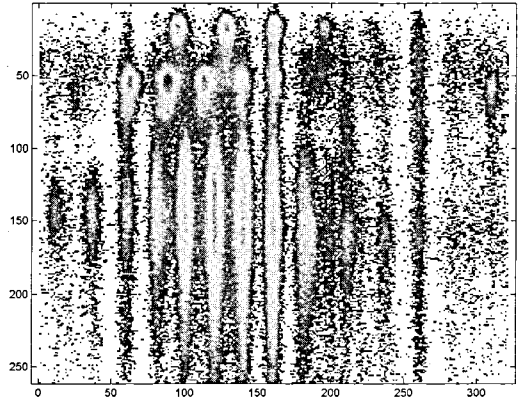
한편, 이온화 방법 중 하나인 ESI(Electro-Spray Ionization)는 분석 물질을 다량의 전하를 띤 이온으로 바꾸게 함으로서, 단백질과 같은 질량이 매우 큰 고분자 화합물의 분석도 가능하게 하였다. 하지만, 이러한 방법은 하나의 이온화된 물질이 다양한 전하 상태를 가지게 함으로서 이온화된 물질로부터 전하 상태를 측정하는 것이 어려워졌다. 또한, 고분자 화합물은 동위 원소를 포함할 확률이 높아 LC/MS 데이터에서 다량의 동위 원소 패턴으로 검출되었다.

본 논문은 과학 기술부의 특정 연구 개발 사업 및 학술진흥재단의 BK21 프로그램으로 지원 받았음

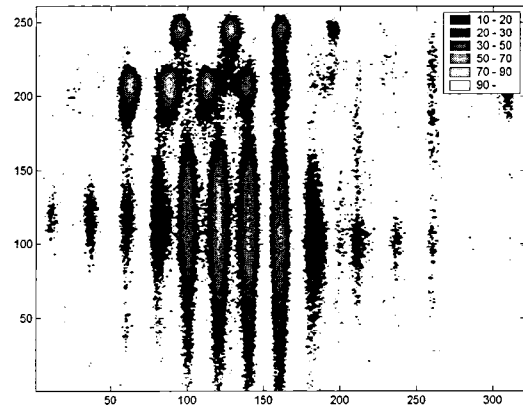
현재까지 질량 스펙트럼과 같은 1차원 신호로부터 디콘볼루션(deconvolution) 과정을 거쳐서 단위 전하당 질량( $m/z$ )값을 순수한 질량값으로 변환시키는 방법들이 많이 연구되어 왔다[1][2][3]. 근래에 LC/MS 장비의 정밀도가 높아짐에 따라 동위 원소 패턴을 이용하여 전하 상태를 측정하는 방법이 연구되었다. Senko et al[4]는 특정 신호 영역에 대하여 패터슨 함수 혹은 푸리에 변환을 통하여 구하여진 전하 맵(charge map)을 이용하였다. 이러한 방법은 SNR이 낮거나 비교적 낮은 수준의 해상도를 가지는 신호에 유용하다. 하지만 하나의 신호영역 마다 오직 하나의 동위 원소 패턴을 찾는다라는 단점이 있다. Z-score 알고리즘[5]은 전하 상태(charge state)를 측정하고 동위 원소 패턴을 찾기 위해 점수 계산 방식을 이용하였다. 하지만 이 방법 역시 복잡한 신호의 경우 쉽게 동위 원소 패턴을 찾기가 힘들었다. MoWeD[6]는 질량 스펙트럼이 아닌 LC/MS 데이터에 대하여 분석을 시도하였다. 하지만, 인접하는 질량 스펙트럼간의 연관성을 살피지 않고 각각의 질량 스펙트럼을 별도로 분석하여 디콘볼루션을 적용하였으며, 후처리 작업으로 인접하는 디콘볼루션된 질량 스펙트럼간의 관계성을 살폈다.

본 논문에서는 질량 스펙트럼을 디콘볼루션할 때, 중요한 단서가 되는 동위 원소 패턴을 찾기 위해서 새로운 접근 방법을 제안하고자 한다. 기존의 접근 방식과 가장 큰 차이점은 LC/MS 데이터를 2차원 영상으로 간주하여  $m/z$  - retention time 도메인에서 직접적으로 분석하는 것이다.

그림 1-1은 실제 LC/MS 데이터의 일부를 가시화(visualization)한 것이다. 가로축은  $m/z$ 값을 세로축은 retention time을 나타낸다.



[그림 1-1] LC/MS를 가시화한 영상. pseudo color로 나타내었으며, 흰색은 intensity가 0인 영역임.



[그림 1-2] 등고선으로 나타낸 영상

그림 1-2는 LC/MS 데이터에서 intensity에 따른 등고선을 나타낸 것이다. 그림을 보면 알 수 있듯이, 시간축으로 연관성이 높고 다양한 길이의 띠 혹은 타원 형태를 이루는 단일 동위 원소 피크들이 존재하며,  $m/z$ 축 방향으로 서로 이웃하는 단일 동위 원소 피크들은 그 형태가 비슷하다. 즉, 같은 동위 원소 패턴에 속하는 단일 동위 원소 피크들은 형태적으로 매우 닮아 있다.

본 논문에서는 이러한 특성에 착안하여 다음과 같은 절차로 동위 원소 패턴을 찾고자 한다. 우선, 전처리 과정으로 LC/MS 데이터에서 노이즈를 제거한다. 그리고 분석 범위를 좁히기 위해, LC/MS 데이터를 각각의 서브 영역으로 분할한다. 다음 과정에서는 각

각의 서브 영역에 대하여 단일 동위 원소 피크를 찾는다. 그리고 검출된 단일 동위 원소 피크로부터 형태적 특징을 추출한다. 마지막으로 이러한 특징을 이용하여 인접한 단일 동위 원소 피크들과의 유사도를 측정하고, 이러한 유사도를 기반으로 점수를 매겨 동위 원소 패턴을 찾는다.

## Materials and Methods

### LC/MS 데이터 획득

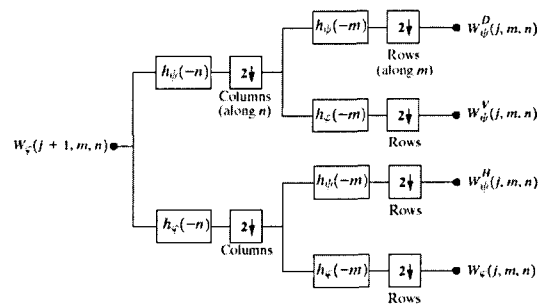
분석 샘플은 펩타이드 라이브러리로서 돼지의 뇌에서 추출하였다. LC/MS 분석 기기는 Ultimate Nano LC System 과 QSTAR PULSAR I Hybrid Q-TOF MS 이며 ESI로 이온화하였다. 그리고 adduct ion은 수소를 사용하였다. LC/MS 데이터는 시간축으로 1초 간격으로 샘플링 되었으며 m/z축으로는 0.01 m/z값 간격으로 샘플링 되었다.

### 노이즈 제거

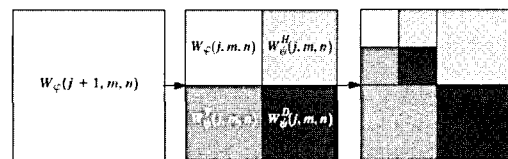
그림1-1 처럼 LC/MS 데이터는 다수의 노이즈를 함유하고 있다. 이러한 노이즈는 신호의 분석을 방해하는 요소이며, 본 연구에서는 크게 세가지 형태의 방해 요소를 정의하였다. 첫번째는 분석 하려는 샘플 자체에서 원래 포함하고 있는 방해 요소이다. 이러한 방해 요소는 신호의 모든 주파수 영역에 영향을 미친다. 이것을 다시 두 가지로 세분화 시키면 저주파 영역에 존재하는 베이스라인 혹은 백그라운드와 고주파 영역에 존재하는 노이즈로 나누어진다. 두번째는 LC/MS 기기를 이용해 샘플을 분리하고 추출할 때, 실험 조건이 나쁘거나 기기 자체의 결함으로 나타나는 신호 왜곡(distortion)이다. 일반적으로 크로마토그래피에서 검출되는 피크는 이상적인 조건이라면 대칭적인

가우시언 형태이어야 하지만, 실제로는 비대칭형으로 나타나는 이유가 바로 이러한 단계에서 발생한다. 마지막 방해요소는 데이터를 얻고 샘플링할 때 발생하는데, 바로 기기 자체의 성능이다. 즉, 정밀도가 떨어지는 기기의 사용으로 인해 저해상도의 데이터를 만들어냄으로써 분석을 어렵게 한다. 본 단계에서는 이러한 방해 요소 중 노이즈를 제거하는데 중점을 두었다. 우리가 찾고자 하는 동위 원소 패턴도 특정한 간격으로 신호가 나타나기 때문에 어느 정도 높은 주파수 성분을 가진다. 따라서, 노이즈를 제거할 때 조심해야 할 필요가 있다. 이러한 상황에 가장 효율적인 방법은 웨이블릿 변환[7][15]을 이용하는 것이다. 즉, 웨이블릿 변환으로 LC/MS 데이터를 특정한 주파수 성분을 가지는 영역으로 분할할 수 있으며, 신호에 영향을 주지 않는 고주파 영역에 대하여 thresholding을 적용한다.

그림 2-1는 2차원 웨이블릿 변환의 절차를 나타낸 것이다. 그림2-2는 다중 해상도 웨이블릿 변환에 대한 도식이다.



[그림2-1] 2차원 이산 웨이블릿 변환



[그림2-2] 다중 해상도 웨이블릿 변환

웨이블릿 필터(wavelet filter)는 널리 쓰이고

있는 Daubechies의 4번째 시리즈[8]를 사용하였으며, 임계값 (threshold value)은 Birge-Massart[9]방법으로 구하였다. 그리고 thresholding 할 때, 수평 방향 디테일과 대각선 방향 디테일은 가장 하위 레벨부터 한 레벨 위까지 적용하였고 수직 방향 디테일은 가장 하위 레벨부터 두 레벨 위까지 적용하였다. 이렇게 차별성을 둔 이유는 m/z 축 방향보다 시간축 방향으로 값들의 변동이 더 심하였기 때문이다.

### 세그멘테이션

노이즈를 제거한 후, 분석 효율을 높이기 위해 각각의 의미 있는 영역으로 분할 (segmentation)하여야 한다. 여기서 의미 있는 영역은 각각의 단일 동위 원소 피크이다. 이러한 서브 영역으로 분할하기 위해 Marker controlled watershed segmentation을 적용하였다. 일반적인 Watershed segmentation [10]은 영역을 과분할(over-segmentation) 시키는 단점이 있다. 따라서, 이를 극복하기 위해 다음과 같은 절차를 수행한다.

### Marker Detection

marker는 인접하는 영역을 구분 짓는 잣대가 된다. LC/MS 데이터에서 좋은 marker는 바로 local maxima가 되는 위치이다. 왜냐하면, 이러한 local maxima가 되는 위치는 단일 동위 원소 피크의 중심에 위치하기 때문이다. 단일 동위 원소 피크는 m/z축과 시간축 사이에 상호 연관성(correlation)이 존재하지 않기 때문에 m/z축 방향으로 local maxima가 되는 지점  $M_x$  과 시간축 방향으로 local maxima가 되는 지점  $M_y$  의 교차점으로 구할 수 있다. 즉, LC/MS 데이터를  $I(x,y)$ , 2차원 가우시언 함수를  $G(x,y)$

라 할 때,

$$M_x = \left\{ (x,y) \left| \begin{array}{l} I(x,y) * \frac{\partial G(x,y)}{\partial x} \leq 0, \\ I(x-1,y) * \frac{\partial G(x,y)}{\partial x} > 0 \end{array} \right. \right\}$$

$$M_y = \left\{ (x,y) \left| \begin{array}{l} I(x,y) * \frac{\partial G(x,y)}{\partial y} \leq 0, \\ I(x,y-1) * \frac{\partial G(x,y)}{\partial y} > 0 \end{array} \right. \right\}$$

가 되며 marker는 집합  $M_x \cap M_y$  에 속하는 공간상의 좌표가 된다.

### Morphological reconstruction using marker

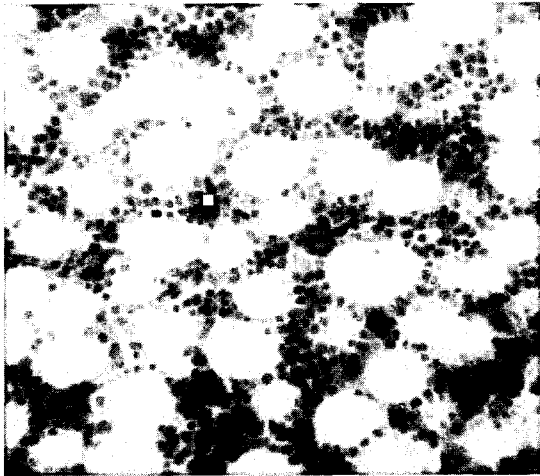
앞에서 찾은 marker를 이용하여 LC/MS 데이터의 intensity를 재구성하여야 한다. 우선, intensity를 음수로 바꾼다. 그러면 marker의 위치는 local minima가 된다. 이제, 이러한 marker를 기준으로 grayscale re-construction [11]을 수행한다. 그림3 은 이러한 과정을 예시로 보여준 것이다. 이러한 reconstruction 과정 후에, 각각의 marker의 위치들만 local minima가 되도록 전체 영상이 재구성되었음을 알 수 있다.

### Watershed segmentation

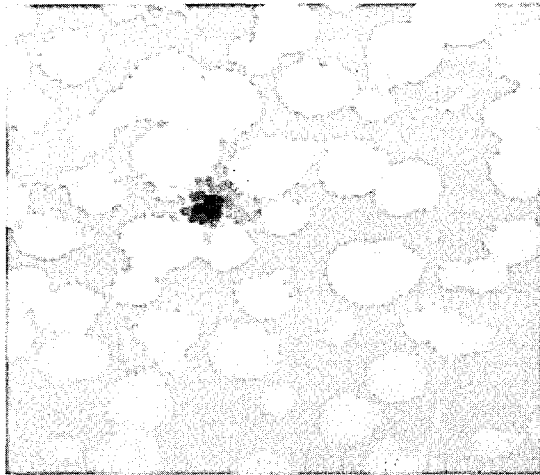
이제 재구성된 데이터에 대하여 전통적인 Watershed segmentation을 적용한다.

### Model based fitting and Feature extraction

이제 각각의 서브 영역에서 단일 동위 원소 피크가 존재하는 지 확인하고, 만약 단일 동위 원소 피크가 존재한다면, 특징 (feature)을 추출해야 한다. 관찰에 의하면, 하나의 단일 동위 원소 피크에 대한 질량 스펙트럼의 경우 어느 정도 일관성이 있어



[그림3-1] 예시 영상과 marker(흰색 사각형)



[그림3-2] Reconstruction 결과

( 대부분이 가우시언 형태이며, 그 넓이도 일정하다. ) 특징으로 사용하기 어려운 반면, 크로마토그램의 경우 그 길이나 모양 등이 다양하게 나왔으며 같은 동위 원소 패턴에 속하는 단일 동위 원소 피크의 경우, 몇 가지 유사한 성질이 존재했다. 이러한 유사한 성질은 바로 peak maximum position  $t_m$ , skewness  $k$ , variance  $\sigma^2$ 이다.

#### Single Isotope Peaks Validation

크로마토그램에서 특징을 추출하기에 앞서, 서브 영역에서 단일 동위 원소 피크의 존재 여부에 대한 테스트를 해야 한다. 이러한 테스트 조건은 다음과 같다.

$$I_{\max} - I_{\min} > T_d$$

$I_{\max}$  는 서브 영역상에 있는 최대 intensity 값이며,  $I_{\min}$  은 최소 intensity 값 그리고  $T_d$  는 두 값의 차이에 대한 임계값이다. 이러한 테스트 조건은 단일 동위 원소 피크의 peakness를 판단하기 위한 것이다.

#### Fitting by EMG Model

통상적으로 크로마토그램에서 peak의 모양을 추정(estimation)하기 위해 여러 가지 parametric model을 적용하여 피팅(fitting)을 하여 왔다.

이상적으로 chromatographic peak는 가우시언 함수로 특징 지을 수 있다. 하지만 실험 과정에서 발생하는 왜곡 현상으로 실제로 비대칭적 가우시언(skewed gaussian)이 되며 널리 알려진 EMG( Exponentially Modified Gaussian ) 함수[12]를 적용하여 피팅 하였다.  $\tau > 0$ 일 때, EMG 함수  $f_{emg}(t)$ 는 다음과 같이 근사화된다.[13]

$$f_{emg}(t) = \frac{M_0}{2\tau} \exp\left(\frac{\sigma_g^2}{2\tau^2} - \frac{t-t_g}{\tau}\right) \times \operatorname{erfc}\left(\frac{1}{\sqrt{2}}\left(\frac{\sigma_g}{\tau} - \frac{t-t_g}{\sigma_g}\right)\right)$$

여기서,  $M_0$  는 피크의 넓이( 0<sup>th</sup> order statistical moment),  $\tau$  는 지수 분포(exponential distribution)의 시간 상수(time constant),  $\sigma_g$  는 가우시언 함수의 표준 편차,  $t_g$  는 가우시언 함수의 retention time이 된다.

각 파라미터값을 추정하기 위해서 Nonlinear optimization 방법 중 하나인 Levenberg-Marquardt Method를 적용하였다. 하지만, 추정되는 파라미터값은 초기값에 큰 영향을 받기 때문에 초기값을 다음과 같이 설정하였다.

$$M_0' = \sum_t P(t)$$

$$t_g' = \arg \max_t P(t)$$

$$\tau' = \frac{1}{\sum_t P(t)} \sum_t tP(t) - t_g'$$

$$\sigma_g'^2 = \frac{1}{\sum_t P(t)} \sum_t t^2 P(t) - \left( \frac{1}{\sum_t P(t)} \sum_t tP(t) \right)^2$$

### Feature Extraction

LM Method에 의해 추정된 파라미터를  $\hat{M}_0, \hat{t}_g, \hat{\tau}, \hat{\sigma}_g$  라 하자. 특징으로 사용되는 peak maximum position  $t_m$ , skewness  $k$ , variance  $\sigma^2$  는 다음과 같다.

$$t_m = t_g', k = \frac{\hat{\tau}}{\hat{\sigma}_g}, \sigma^2 = \hat{\sigma}_g^2$$

그리고 이러한 특징들을 이용하여 다음과 같은 특징벡터  $\mathbf{f}$  를 얻을 수 있다.

$$\mathbf{f} = \begin{bmatrix} t_m \\ k \\ \sigma^2 \end{bmatrix}$$

### Isotope pattern grouping

각각의 단일 동위 원소 피크로부터 얻은 특징 벡터  $\mathbf{f}$  를 이용하여 인접한 단일 동위 원소 피크끼리 유사도를 비교하고 그루핑하여야 한다.

우리는 z-score 알고리즘과 비슷한 방법을 이용하였다. 즉, 각각의 후보 전하값마다 점수를 계산하고 가장 점수가 높은 전하값에 해당되는 단일 동위 원소 피크들을 묶는 것이다. 유사도 측정을 위해 z-score 알고리즘이 intensity 정보만 이용했으나 보다 풍부한 정보인 특징 벡터  $\mathbf{f}$  를 이용하여 유사도를 측정하였다.

$\mathbf{F}$  를 벡터  $\mathbf{f}$  들의 집합이라고 하자. 이 때,

그루핑(grouping) 알고리즘은 표1과 같다.

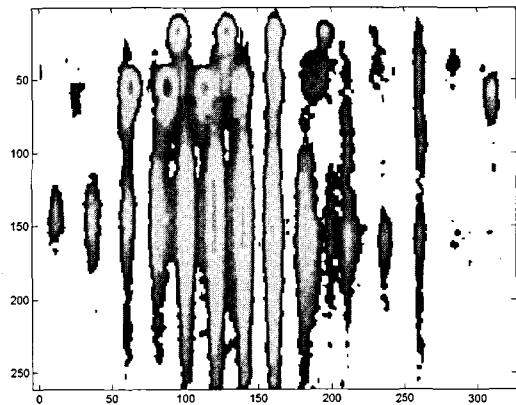
```

while  $\mathbf{F}$  is not empty
  find  $f_i$  whose intensity of  $t_m$  is maximum
  for each assumed charge state  $z_i$ 
    find other expected single isotope peak  $f_j$ 
    depending on the gap
    calculate  $\mu_1, \mu_2, \mu_3, \sigma_1^2, \sigma_2^2, \sigma_3^2$ 
    each of which is sample mean and sample variance
    of features  $t_m, k, \sigma^2$ 
    crop  $f_j$  in which any feature is out of
      interval  $(\mu_i - 1.96\sigma_i, \mu_i + 1.96\sigma_i), i = 1..3$ 
     $\text{score}(f_i; z_i) = -\frac{\sigma_1^2 + \sigma_2^2 + \sigma_3^2}{n}$ 
    where  $n$  is number of inliers  $f_j$ 
    group single isotope peaks  $f_i, \hat{f}_j$ 
    such that  $\arg \max_{i,j} \text{score}(f_i; z_i)$ 
  remove  $f_i, \hat{f}_j$  from  $\mathbf{F}$ 
  
```

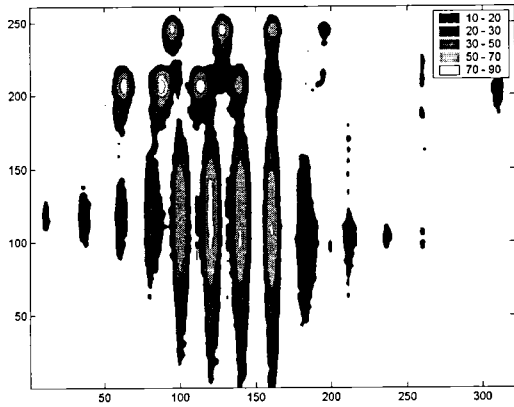
[표1] 그루핑 알고리즘

### Experiment and Results

그림 3-1은 그림1-1의 데이터로부터 노이즈를 제거한 영상이다. 그림 3-2는 노이즈가 제거된 영상을 등고선으로 나타낸 영상이다.



[그림 3-1] 노이즈가 제거된 LC/MS 영상



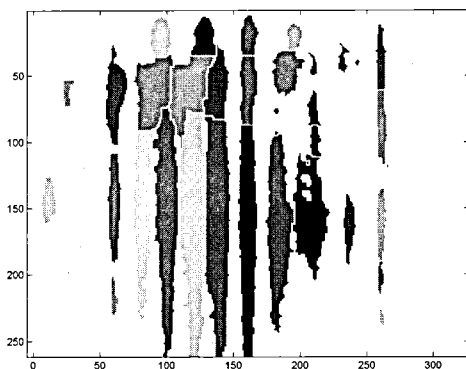
[그림 3-2] 노이즈가 제거된 영상의 등고선 그림

육안으로 확인할 수 있듯이 원본 데이터에 비하여 랜덤 노이즈가 많이 제거되었으며, 등고선의 볼록함( convexness )도 훨씬 좋아졌음을 알 수 있다.

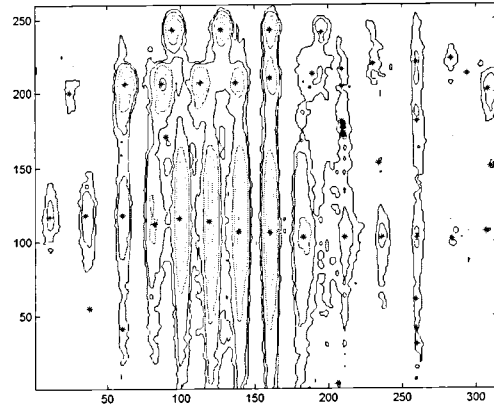
그림 4-1은 영상 분할 결과를 보여준다. 하나의 영역은 동일한 색으로 칠해져 있으며 이러한 영역에는 오직 하나의 local maxima가 존재한다. 그 결과 단일 동위 원소 피크가 존재하는 영역으로 잘 분할되어 있다.

그림 4-2는 marker로서 검출된 local maxima의 위치를 나타낸다. 일부는 local maxima의 위치가 잘못되었지만, 대부분의 local maxima의 위치가 제대로 찾아졌음을 알 수 있다.

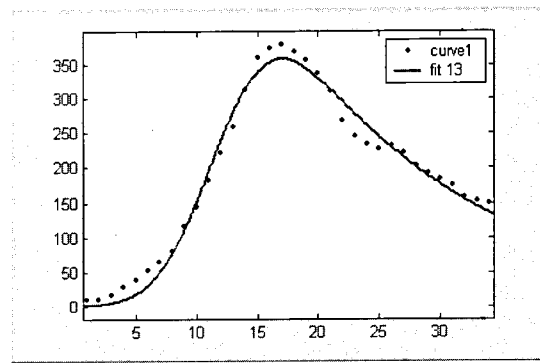
그림 5는 그림 4-1에서 특정 단일 동위 원소 피크 영역( 화살표로 표시 )의 크로마토



[그림 4-1] marker controlled watershed segmentation결과



[그림 4-2] local maxima의 위치



[그림 5] 실제 데이터와 피팅된 모델.

(점 - 추출된 크로마토그램, 선- 피팅된 모델) 그램과 EMG 모델을 이용하여 피팅시킨 결과이다. 비대칭성(asymmetry)을 고려하여 피팅이 되었음을 알 수 있다.

파라미터	초기값	LM 방법으로 추정된 값
$M_0$	6661.4	8562
$t_g$	17	11.84
$\tau$	3.0125	15.22
$\sigma_g$	7.1505	3.79

[표 2-1] 그림 5 크로마토그램의 파라미터의 초기값과 추정된 값

특징	$t_m$	$k$	$\sigma$
값	17	4.01	3.79

[표 2-2] 계산된 특징들

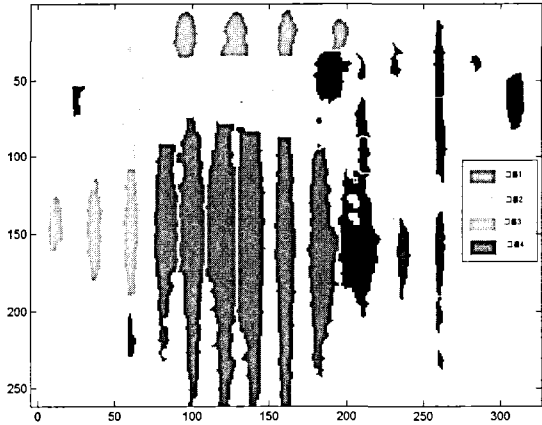
표2-1은 그림 5의 피팅된 EMG 함수에서 추정된 파라미터를 나타내고 있으며 표2-2는 이로부터 얻어지는 특징을 나타낸다. LM 방법과 같은 비선형 방법은 초기값이 중요하여 이러한 초기값들은 EMG 함수의 통계학적 특성으로부터 근사화시킬 수 있다 [14].

그림 6은 표1의 알고리즘을 이용하여, 그림 4-1에서 세그멘테이션 된 영역에서 추출된 특징을 기반으로 특징값들이 유사한 단일 동위 원소 피크끼리 그루핑한 결과이다. 표 3은 그루핑된 동위 원소 패턴들의 점수를 보여준다. 검정색 영역은 다른 단일 동위 원소 피크들과의 유사성을 찾지 못한 영역이다. 그 이유는 영상 분할이 제대로 수행되지 못해서 두 개의 단일 동위 원소 피크가 합쳐져 있거나 분할된 영역이 크로마토그램을 표현하기에는 데이터의 양이 작아서 unstable하기 때문이다.

표 3은 그루핑된 단일 동위 원소 패턴들의 특징 벡터들과 점수값을 나타낸 것이다. 음수값으로 나오는 이유는 가장 큰 값을 점수값으로 결정하기 위해서이다. 점수 계산을 위해서 분산값을 이용하였기 때문에 분산값들이 작을수록 같은 그룹에 속하는 동위 원소 패턴이 될 확률이 높다.

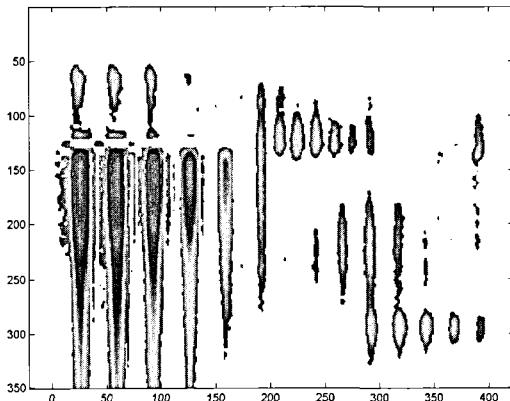
그룹명	점수
그룹1	-1.5739
그룹2	-2.4680
그룹3	-4.6788
그룹4	-4.3175

[표 3] 각 동위 원소 패턴들의 스코어



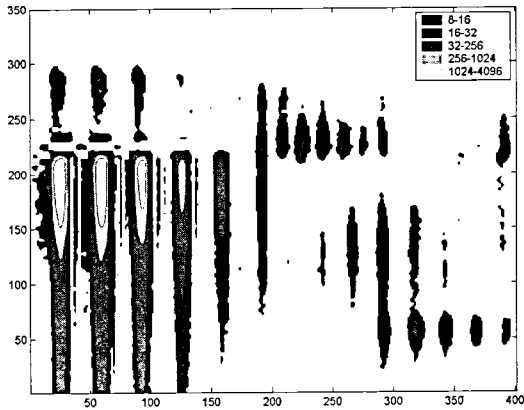
[그림 6] 특징들을 기반으로 그루핑한 결과

다음은 하나의 동위 원소 패턴에 속하는 단일 동위 원소 피크들의 특징들이 얼마나 유사한지 살펴보도록 하자. 그림 7-1은 다양한 형태의 동위 원소 패턴들이 나타나 있다. 그림 7-2는 이러한 동위 원소 패턴들의 등고선을 나타낸 것이다. 검출되는 intensity값은 최대 약 4000까지 치솟는다.



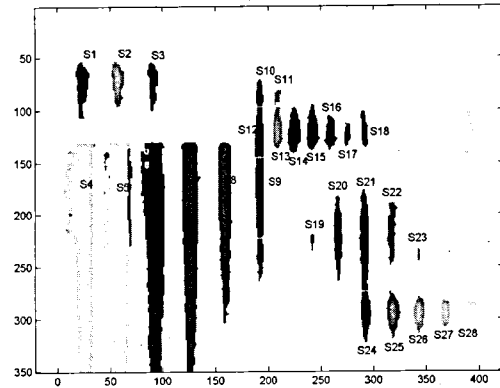
[그림7-1] 다양한 동위 원소 패턴들이 나타나 있는 LC/MS 데이터





[그림 7-2] 등고선 그림( intensity값이 매우 크게 변함을 알 수 있다.)

그림 7-3은 각 동위 원소 패턴에 라벨(label)을 붙인 것이다. 표 4는 각각의 단일 동위 원소 패턴에서 추출된 특징들을 보여준다.  $P_x, P_y$  는 peak maximum이 되는 위치이다. skewness  $k$  값이 음수로 나오는 것은 크로마토그램에서 peak maxima position이 오른쪽으로 치우친 형태가 되며 반대로 양수이면 왼쪽으로 치우친 형태가 된다. SIP는 단일 동위 원소를 의미하며 (\*)로 표시된 것은 그룹핑 과정에서 outlier로 판명난 것이다. 표에서 알 수 있듯이 같은 그룹에 속하는 각 특징값들은 비슷한 값을 가지며, 연관성이 존재한다. 예를 들어, skewness 값은 같은 그룹에 속할 때, 같은 부호를 가진다. 또한, 피팅이 안되는 단일 동위 원소 피크들이 존재하는데, 피팅이 안되는 이유는 단일 동위 원소 피크의 영역이 넓지 못해서 충분한 양의 데이터를 뽑지 못했기 때문이다. 그룹 E에 피팅이 제대로 안 되는 피크들이 많은 이유가 그림에서 알 수 있듯이 피크들의 넓이가 작기 때문이다. R-square는 피팅 에러를 나타내며 charge state는 동위 원소 패턴의 계산된 전하값이다.



[그림 7-3]

## Discussion

본 논문에서는 LC/MS 데이터에서 동위 원소 패턴을 찾기 위해 2차원 공간 상에서 접근하는 방법에 대해 살펴보았다. LC/MS 데이터를 가시화 해보면, 기존의 1차원 신호로 모델링하였던 동위 원소 패턴은 2차원 영상으로 모델링할 때, 더 많은 정보가 나타남을 알 수 있다. 즉, 하나의 동위 원소 패턴에 속하는 각각의 동위 원소 피크들은 peak maximum position 뿐만이 아니라 skewness, variance등 유사점이 많이 나타났다. 따라서 동위 원소 패턴을 찾을 때, 이러한 추가적인 특징을 이용할 경우 보다 더 정확하게 찾을 수 있다. 하지만, 실제로 서로 다른 동위 원소 패턴에 속하는 단일 동위 원소 피크들이 비슷한 공간에 겹쳐서 나타나는 경우가 많기 때문에 영상 분할 단계에서 단일 동위 원소 피크들을 분리하기가 힘들다. 따라서, 보다 견고한 분석을 위해서 top-down 접근 방식을 도입하여야 한다. 즉, 같은 동위 원소 패턴에 속하는 동위 원소 피크들의 경우 본 논문에서 살펴 본 바에 의하면 2차원적 특징이 비슷하기 때문에 동위 원소 패턴 자체에 대한 parametric model을 세울 수 있다. 그리고 2차원 공간상에서 다양한 동위 원소 패턴들의 mixture 함수로

서 LC/MS 데이터를 정의할 수 있다. 앞으로 이러한 모델을 세우고 여러 동위 원소 패턴들이 합쳐져서 나타날 경우에 분리시키는 방법에 대해서 연구를 할 계획이다.

## References

- [1] M. Mann, C.K. Meng and John B. Fenn, Interpreting Mass Spectra of Multiply Charged Ions, *Rapid Commun. Mass Spectrom*, 7, 71-54
- [2] Bruce R. Reinhold and Vernon N. Reinhold, Electrospray Ionization Mass Spectrometry : Deconvolution by an Entropy -Based Algorithm, *J. American Society for Mass Spectrometry*, 3, 207-215
- [3] Jefferey J. Hagen and Curtis A. Monnig, Method for Estimating Molecular Mass from Electrospray Spectra, *Anal. Chem*, 66, 1877-1883
- [4] Michael W. Senko et al. Automated Assignment of Charge State from Resolved Isotopic Peaks for Multiply Charged Ions, *J. American Society for Mass Spectrometry*, 6, 52-56
- [5] Zhongqi Zhang and Alan G. Marshall, A Universal Algorithm for Fast and Automated Charge State Deconvolution, *J. American Society for Mass Spectrometry*, 11, 329-332
- [6] John O. Percy and Terry D. Lee, MoWed, a Computer Program to Rapidly Deconvolute Low Resolution Electrospray Liquid Chromatography /Mass Spectrometry Runs to Determine Component Molecular Weights, *J. American Society for Mass Spectrometry*, 12, 599-606
- [7] Rafael C. Gonzalez, Digital Image Processing, 2002, Prentice Hall
- [8] Daubechies, I., Ten lectures on wavelets, CBMS-NSF conference series in applied mathematics, SIAM Ed.
- [9] Birgé, L. and P. Massart, From model selection to adaptive estimation, D. Pollard (ed), Festschrift for L. Le Cam, Springer, pp. 55-88.
- [10] F. Meyer and S. Beucher, Morphological segmentation, *Journal of Visual Communication and Image Representation*, 1 (1), 21-46, September 1990.
- [11] Luc Vincent, Morphological Grayscale Reconstruction in Image Analysis: Applications and Efficient Algorithms, *IEEE Transactions on Image Processing*, 2(2), April 1993, pp. 176-201
- [12] A.N. Papas and T.P. Tougas, Accuracy of Peak Deconvolution Algorithms within Chromatographic integrators, *Anal. Chem*, 62, 234-239
- [13] D. Hanggi, P.W. Carr, Errors in exponentially modified Gaussian equations in the literature, *Anal. Chem*, 57, 2394-2395
- [14] E. Grushka, Characterization of exponentially modified Gaussian peaks in chromatography, *Anal. Chem*, 44, 1733-1738
- [15] M. Vetterli, J. Kovacevic, Wavelets and Subband Coding, Prentice Hall, 1995

Group	SIP	$P_x$	$P_y(T_m)$	$K(\tau/\sigma)$	$\sigma$	R-square	charge state
Group A	S1	23	68	4.237	3.415	0.9712	3
	S2	57	67	3.9	3.876	0.9759	
	S3	90	67	6.29	3.177	0.9552	
Group B	S4	24	152	5.23	6.746	0.9820	3
	S5	58	153	5.04	7.397	0.9804	
	S6	92	152	3.12	11.63	0.9891	
	S7	125	152	6.79	5.984	0.9926	
	S8	158	153	6.31	8.536	0.9799	
	S9(*)	191	162	0.58	43.52	0.9495	
Group C	S10	191	88	-1.84	5.343	0.9434	6
	S11	208	88	-2.07	5.193	0.8384	
Group D	S12(*)	191	131	-0.78	19.48	0.9614	6
	S13	208	123	-3.39	4.897	0.9421	
	S14	224	124	-2.47	6.355	0.9784	
	S15	241	123	-2.24	5.908	0.9663	
	S16	258	124	-5.36	3.199	0.9603	
	S17	274	124	-5.52	3.578	0.8908	
	S18	290	122	-5.35	5.191	0.8882	
Group E	S19	241	225	Not fit			4
	S20	265	222	0.4	20.57	0.9459	
	S21	290	226	1.7	20.88	0.9285	
	S22	316	216	Not fit			
	S23	341	217	Not fit			
Group F	S24	291	294	0.72	5.875	0.9776	4
	S25	317	294	0.8	7.415	0.9904	
	S26	342	295	0.8	6.909	0.9770	
	S27	367	295	0.86	6.378	0.9552	
	S28	392	295	Not fit	15.51	0.8704	

[표 4] 동위 원소 패턴에 속한 단일 동위 원소 피크들의 추정된 특징들