단백질 질량분석을 위한 THRASH 알고리즘 속도 향상 기법[1)]

전상현[O*] 장형수* 오한빈**
*서강대학교 컴퓨터학과, **서강대학교 화학과
{shjeon[O*], hschang*}@smolab.sogang.ac.kr, {hanbinoh**}@sogang.ac.kr

Sogang University

Sang-hyun Jeon[O*] Hyeong Soo Chang* Han Bin Oh**
*Department of Computer Science, Sogang University
**Department of Chemistry, Sogang University
*,**Program of Integrated Biotechnology, Sogang University

요 약

Horn *et al.*에 의하여 개발된 THRASH 알고리즘은 대표적인 단백질 질량분석 플랫폼으로써, 극초분해능(ultra high resolution) Fourier Transform 질량분석법을 통해 얻어지는 고집적 전기분무 (electrospray ionization, ESI) 질량 스펙트럼 데이터를 분석하는데 이용되고 있다. 하지만 이 알고리즘은 속도 면에서 부족하여 실시간 분석에 한계점을 보이고 있다. 이를 보완하기 위해 본 논문에서는 THRASH알고리즘의 속도를 향상시키는 기법을 제안하고 실험 결과를 통하여 새로운 기법이 융합된 알고리즘의 수행 속도가 기존 THRASH알고리즘의 속도를 비약적으로 향상 시킬 수 있음을 보인다.

1. 서론

본 논문에서는 Horn *et al.* [4] 에 의하여 개발된 전기분무 (electrospray ionization: ESI) [1] 질량 스펙트럼(ESI-mass spectrum) 데이터를 분석하는 기존의 THRASH 알고리즘을 분석해 속도 면에서 보다 향상된 알고리즘을 제안하고 이를 실험적으로 구현, 확인하려 한다.

최근에 휴먼게놈 프로젝트가 성공적으로 수행됨에 따라 생명체에 대한 게놈 수준의 정보가 전례 없이 증가하였다. 하지만, 활발한 생명과학/공학 분야의 연구들을 통하여 게놈 수준의 이해를 초월한 생명정보의 생체 내 구현형태인 단백질체(Proteome) 수준에서의 이해가 필수적이라는 점을 인식하게 되었다. 이러한 배경 하에 소위 단백질체학(Proteomics) 이라고 불리우는 새로운 학문 분야를 탄생케 하였다. 이 단백질체학에서 주요한 실험적 도구로써 질량분석법이 자리를 매김하고 있는데, 특히 가장 고질량분해능을 자랑하는 Fourier Transform(FT) 질량분석법이 최근에 들어 많은 주목을 받고 있다. 하지만, 이 FT 질량분석기에서 얻어지는 고집적(data intensive) 질량분석스펙트럼을 분석하는 소프트웨어의 개발 및 활용 수준은 비교적 낮은 분해능을 가진 저집적 질량분석스펙트럼 분석에 비해 초기단계에 지나지 않는다. 특히 단백질체학 및 바이오인포매틱스 분야에서 해결해야 할 주목받는 과제로써 고속 질량분석 및 해석을 들 수 있다.

자연계에는 수많은 동위원소가 존재하기 때문에 어떤 분자의 질량을 단일한 것으로 생각할 수 없다. 즉 같은 종류인 분자들의 질량은 isotopic cluster(a set of isotopic peaks from a single molecular or fragmentation entity)를 형성한다. isotopic cluster는 분자에 따라 고유한 분포를 가지므로 질량분석을 통해 분자의 종류를 대략적으로 판별할 수 있다. 단백질 분자와 같은 고분자의 질량측정 시 ESI 방법에 의해 높은 charge수를(z) 갖기 때문에, 질량 값에 비해 작은 m/z(mass per charge)값으로 표현된다. 이러한 이유로 인해 고분해능 질량분석 스펙트럼의 해석을 위해서는 고유한 알고리즘이 필요하다. 현재로써는 ESI 질량 스펙트럼 데이터를 분석하기 위해 개발된 알고리즘들 중 THRASH 알고리즘이 "de facto standard"로서 자리매김하고 있다[2][3]. 하지만 현재의 THRASH 알고리즘의 속도는 이러한 필요성을 충족시키지 못하고 있다. 본 논문에서는 새로운 THRASH 알고리즘의 속도 향상 기법을 적용해 THRASH의 정확성은 거의 손상시키지 않으면서 최대 4배의 속도 향상을 이룰 수 있다는 사실을 확인하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 THRASH 알고리즘의 동작과정을 간략히 소개하며, 3, 4장에서는 현재 THRASH알고리즘 상에서 문제점을 분석해 이러한 문제점을 보완하는 방안 및 속도향상 결과를 각기 제시하고자 한다.

2. THRASH 알고리즘 동작과정

THRASH 알고리즘의 주요 목적은 입력 ESI 질량 스펙트럼 내에서 isotopic cluster들을 찾는 것이다. THRASH알고리즘의 전체적인 흐름은 아래 [그림 1]과 같다.

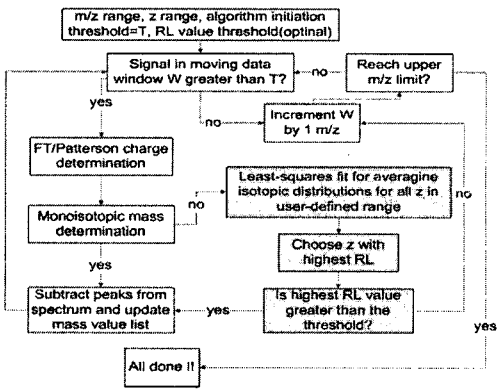

그림 1. THRASH 알고리즘의 흐름도[4]

위의 [그림 1]에서 isotopic cluster를 찾는 과정을 간단히 정리하면 다음과 같다 (논문[4]를 참조).

[Find signal] 1-m/z data window구간 내에 있는 T보다 높

1) 본 연구는 한국기초과학지원연구원의 "다목적 첨단 질량분석장치 개발사업" 으로부터 연구비 지원을 받았습니다. 본 연구를 위해 많은 조언을 해 주신 한국기초과학지원연구원의 유종신, 김현식, 권경훈 박사님, 허만회 연구원께 큰 사의를 드립니다.

은 peak 중에서 most abundant peak을 찾는다.

[Charge determination] FT/Patterson 방법을 이용해 찾아낸 peak의 most probable charge를 찾는다.

[Calculate theoretical distribution] 다음으로 이 charge값을 이용해 most abundant peak의 질량을 구한다([7] 참조). 이렇게 구한 peak의 질량을 이용해 현재 peak가 가질 수 있는 theoretical isotopic abundance distribution을 구한다.

[Match distribution] 이렇게 구한 theoretical isotopic abundance distribution과 입력 질량 스펙트럼을 least-square fitting 방법을 이용해 두 개의 peak분포가 얼마나 유사한가를 나타내는 *Figure of Merit* (FOM)값을 구한다. FOM값을 이용해 적합도 *RL value* (reliability value)를 구한다.

[Match distribution:success]앞에서 구한 RL value가 90% 이상이라면 올바른 isotopic cluster를 찾은 것으로 간주해 입력 spectrum에서 찾아낸 theoretical distribution을 제거한 뒤에 계속해서 다른 isotopic cluster를 찾는 작업을 계속한다.

[Match distribution:fail] 한편, RL value가 90% 이하라면 현재 most abundant peak가 가질 수 있는 모든 charge값에 대해서 theoretical distribution을 구한 뒤에 least-square fitting을 이용해 각 distribution에 대한 RL value를 구한다. 이 중 가장 높은 RL value가 90%를 넘는다면 peak에 해당하는 charge값을 할당해주고 theoretical distribution을 제거한 뒤에 다른 isotopic cluster를 찾는 작업을 계속한다.

[Increase data window] 현재 data window구간 내에 threshold보다 높은 peak이 없다면 data window구간을 1-m/z 증가 시킨 뒤 위의 과정을 반복한다.

다음 장에서는 현재 THRASH알고리즘 상에서 문제점을 분석해 이러한 문제점을 보완하는 방안에 대해서 서술한다.

3. THRASH 알고리즘의 분석 및 성능 향상 기법

Charge determination 부분에서 하는 수행하는 일은 특정 peak의 charge를(z) 결정하는 것이다. 기본적으로 charge를 결정하는 원리는 다음과 같다. 자연계에 존재하는 동위원소의 분포 특성상 ^{12}C, ^{13}C가 전체 질량에 기여하는 비중이 가장 크다[7]. 따라서 가로축이 질량(m)인 스펙트럼에서 각 isotope peak 사이의 거리는 약 1Da([4][7]을 참조)가 나기 때문에 가로축이 m/z일 경우 1 m/z구간 내에 있는 peak의 개수를 알면 charge를 알아낼 수 있다는 사실과 charge값이 정수 값만을 가진다는 사실에 기초하고 있다[7]. 따라서 FT를 이용하면 각 frequency 마다 abundance를 구할 수 있기 때문에 charge를 쉽게 구할 수 있다. THRASH알고리즘에서 사용하는 방법은 보다 정확도가 향상된 기법으로 FT와 Patterson 방법을 혼합한 방법을 사용한다(자세한 설명은 논문[6]을 참조).

다음으로 입력 스펙트럼과 theoretical isotopic distribution을 matching하기 위해서 THRASH에서는 least-square fitting방법을 이용한다. 이 때 사용되는 FOM값은 다음과 같은 수식을 통해 얻어진다 :

$$FOM = C/\sum_n [(A_n - I_n)^2 - (NV)^2]$$

A_n: 이론적인 abundance distribution에서 n번째 peak의 상대적인 abundance

I_n : n번째 peak의 m/z값에서의 spectral intensity

V : 인접한 valley에서의 최대 spectral intensity

N : 정규화 계수, C: 비교횟수

이제 THRASH 알고리즘이 가지는 수행속도 면에서의 문제점을 분석하기로 한다. Bovine ubiquitin, 8.6kDa, 76 amino acid의 ESI 질량 스펙트럼을 테스트 데이터로 사용한 결과 다음과 같은 사실을 알 수 있었다.

우선 THRASH알고리즘의 주요 구성요소가 전체 수행시간에 차지하는 비율을 측정해 본 결과 least-square fitting이 전체 수행시간의 95% 이상을 차지하고, charge determination부분이 약 2%를 차지하는 것을 알 수 있었다. 따라서 least-square fitting과 관련된 부분의 최적화에 중점을 두었다. 다음으로 THRASH알고리즘의 수행과정 중 least-square fitting이 실패했을 경우, 즉 RL value 가 90% 이하일 경우 수행하는 부분(그림 2에서 노란색으로 표시된 부분. 이 부분을 앞으로 *backup method*라고 언급한다.)을 알고리즘에서 제외한 경우와 그렇지 않은 경우 두 방법의 수행속도와 찾아낸 charge개수에 대한 비교를 해 본 결과 *backup method*를 사용하지 않을 경우, 수행시간이 10배 정도 빨라지고 찾아내는 peak 개수는 75개로 기존 THRASH알고리즘이 찾아낸 83개에 비해 단지 8개만 적은 peak를 찾아냈다는 사실을 확인했다. 즉 THRASH알고리즘이 전체 83개의 peak중에서 8개의 peak을 찾기 위해 알고리즘 전체 수행시간의 약 90%가 낭비되고 있다는 것이다. 이는 THRASH알고리즘에서 할당 가능한 모든 charge에 대해서 least-square fitting을 수행함에 따라서, 실제 알고리즘이 수행할 때 peak의 charge가 될 가능성이 없거나 아주 낮은 charge에 대해서도 모두 least-square fitting을 수행하기 때문에 이러한 결과가 나왔다.

앞에서 살펴본 바와 같이 *backup method*부분에서 무의미한 search를 많이 하므로 이러한 search의 횟수를 줄이면 전체 알고리즘의 성능을 향상시킬 수 있다. 실제로 THRASH 알고리즘에서는 *backup method* 부분에서 실행 가능한 모든 charge에 대해서 least-square fitting을 수행하고 있는데 그럴 필요는 없다. 이 경우 만약 RL value가 90%를 넘는 charge가 있는 경우에는 상관없지만 그렇지 않을 경우에는, fitting을 수행하는 charge 개수만큼 전체 알고리즘의 성능에 부담을 준다. 실제로 알고리즘 수행도중 *backup method*를 수행하는 횟수와 실제로 이 부분에서 찾아내는 charge의 개수를 비교해보면 위의 실험 데이터의 경우 총 168번 *backup method*가 수행되었는데 이 중 실제로 찾아낸 peak의 개수는 8개에 불과하다. 따라서 *backup method*에서 검색하는 charge의 개수를 줄이면 전체적인 THRASH알고리즘 수행시간을 많이 줄일 수 있다는 것을 알 수 있다.

따라서 각 charge에 대해서 least-square fitting을 수행할 때 실제 charge가 될 가능성이 높은 *charge set*에 대해서만 fitting을 수행한다. 이러한 *set*은 다음과 같이 결정한다.

[Add at most *n* charges]. FT/Patterson 방법으로 나온 각 charge값에 대한 점수 중 가장 높은 순서대로 *n*개(n ≤ maximum charge)의 charge를 charge set에 넣는다. 실제로 charge determination 부분에서 FT/Patterson방법이 수행된 후의 결과는 continuous값이다. 하지만 charge는 discrete값이므로 변환과정에서 오차가 생기기도 하고, 또한 서로 배수관계에 있는 charge의 경우(3, 6, 9 같은 경우)에는 charge 3의 peak을 charge 6, 9가 공유하기 때문에 이러한 과정에서 오차가 생기기도 한다.

[Add charge 1, 2]. charge 1, 2를 charge set에 넣는다. 이는 THRASH에서 1-m/z data window를 사용하기 때문에 charge 1, 2는 인식하기가 쉽지 않다. 그 이유는 charge가 1(2)인 경우는 data window내에 peak이 1(2)개만 있어야 하므로 이를 정확히 찾아내기가 쉽지 않다 (논문[6]을 참조).

이와 같이 *charge set*을 설정한 뒤에 *backup method*부분에서 *charge set*에 대해서만 fitting을 시도하게 되면 전체적으로 알고리즘의 수행시간이 많이 줄어드는 것을 알 수 있다. 자세한 결과는 다음 장에서 다룬다.

4. 성능평가

4. 1. 실험 환경 및 평가

실험에 사용된 컴퓨터 사양은 Intel Pentium 4 3.2G CPU, 512MB이다. 실행 프로그램은 윈도우 환경에서 컴파일 한 뒤에 실행했다. 입력 질량 스펙트럼은 Bovine ubiquitin, 8.6kDa, 76 amino acid로부터 나온 293개의 데이터들을 이용했다. 각각의 데이터들은 같은 단백질로부터 얻어졌지만 ESI 질량 스펙트럼을 m/z구간별로 자른 것이기 때문에 서로 독립적인 데이터로 볼 수 있다. THRASH알고리즘에서 설정한 maximum charge 값은 20이고, 향상된 알고리즘의 charge set 개수는 4, 5, 6, 7, 8, 10개로 정했다(여기서 maximum charge값은 THRASH에서 charge determination이나 *backup method*가 수행될 때 검사하는 최대 charge값을 의미한다). 먼저 각 입력 데이터를 기존의 THRASH 알고리즘으로 수행시킨 결과와 각각의 *charge set*의 개수에 따른 수행결과를 비교했다. 비교항목은 각각의 입력 데이터에 대한 속도와 찾아낸 peak수가 된다.

4. 2. 실험 결과 및 분석

[그림 2]는 총 입력 데이터에 대한 각 방법의 수행속도를 THRASH알고리즘의 속도를 기준으로 정렬한 데이터를 그래프로 표시한 것이다. 여기서 가로축은 각각의 입력 데이터를 의미하고 세로축은 해당 데이터에 대한 각 방법의 수행속도를 의미한다. 다음으로 peak error를 THRASH알고리즘에서 찾아낸 peak수에서 향상된 알고리즘에서 찾아낸 peak수를 뺀 값으로 정의하자. 다음 [표 1]에는 charge set의 개수가 5일 경우 peak error값에 따른 입력 데이터의 개수를 정리했다. 다른 charge set의 개수에 대해서도 유사한 결과를 나타내기 때문에 여기서는 제외했다. [표 2]에는 각각의 *charge set*의 개수에 따른 전체 입력 데이터에 대한 평균 수행시간과 peak error가 0이 아닌 입력데이터의 개수를 정리했다.

[그림 2]를 통해 알 수 있듯이 각 입력 데이터에 대한 수행속도는 *charge set*의 개수가 작을수록 빨라지는 것을 알 수 있다. [표 1]에는 charge set의 개수가 5인 경우 peak error를 보여주고 있는데 94.5%의 경우 기존의 THRASH프로그램과 같은 결과를 보여주었고 나머지 경우에만 찾아낸 peak의 개수가 1~2개 차이를 보였다. 하지만 이와 같이 peak의 개수가 차이를 보이는 경우에도 나머지 peak들은 기존THRASH알고리즘이 찾아낸 peak와 같다는 것을 확인했다. 다음으로 [표 2]를 보면 charge set의 개수와 이에 따른 peak error를 나타내고 있다. [표 2]를 통해 알 수 있듯이 *charge set*의 개수가 많아질수록 peak error는 줄어드는 반면 수행속도는 늘어나는 것을 알 수 있다. 표에는 나와 있지 않지만 모든 경우를 통틀어 peak error 값의 오차가 ±3개 이내인 것을 알 수 있었는데 이는 향상된 알고리즘이 속도뿐만 아니라 정확도 면에서도 상당히 좋은 결과를 보이는 것을 확인할 수 있다. 위 실험 결과에 따르면 수행속도와 peak error의 적절한 균형을 맞추기 위해서 charge set의 개수를 5-7로 하는 것이 적절하다. charge set의 개수가 5인 경우 기존 THRASH알고리즘과 향상된 알고리즘의 속도를 비교해 보니 1.8-4.3배 정도의 차이를 보였다. 한편 charge set의 개수가 8인 경우와 10인 경우 peak error가 같게 나오는데, 이처럼 *charge set*의 개수가 늘어나도 정확도가 향상되지 않는 경우에 대해서는 추후 연구가 필요하다고 본다.

5. 결론

지금까지 살펴본 바와 같이 단백질 질량 분석을 위해 널리 사용되고 있는 THRASH알고리즘의 향상 방안에 대해서 알아보았다. 앞에서 살펴본 바와 같이 *charge set*의 개수를 조정함에 따라 실제 수행속도 면에서는 많은 성능향상을 가져왔음을 알 수 있다. 향상된 알고리즘이 찾아내는 peak의 수도 거의 정확한 결과를 보임을 알 수 있었다.

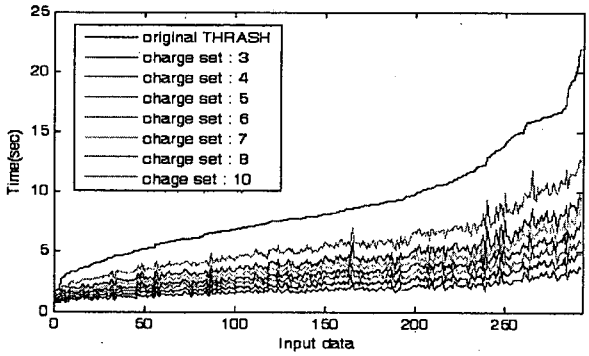

그림 2. 각 charge set의 개수에 따른 각각의 입력 데이터에 대한 수행속도

peak error	데이터 개수
0	277
+1	13
+2	1
-1	2

표 1. charge set의 개수가 5일 때 각 peak error에 따른 입력 데이터의 개수(총 데이터의 개수는 293개)

수행 알고리즘		평균 수행시간(초)	peak error가 0이 아닌 데이터 개수(%)
THRASH		8.974445	0(0 %)
charge set의 개수	3	1.890729	43(14.6%)
	4	2.370445	26(8.9%)
	5	2.876918	16(5.5%)
	6	3.329829	14(4.8%)
	7	3.789507	12(4.1%)
	8	4.443644	11(3.8%)
	10	5.735253	11(3.8%)

표 2. 각 charge set의 개수에 따른 평균 수행속도와 peak error가 0이 아닌 데이터의 개수(총 데이터의 개수는 293개)

참 고 문 헌

[1] Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. "Electrospray ionization-principles and practice," *Mass Spectrom. Rev. 9*, 37-70, 1990.

[2] Ruedi Aebersold; Matthias Mann "Mass spectrometry-based proteomics," Nature publicshing group. 198-207, 2003.

[3] Martha M. Vestling "Using mass spectrometry for proteins," Journal of chemical education vol. 80 no. 2 122-124, February 2003.

[4] Horn, D. M., Zubarev, R. A., McLafferty, F. W., "Automated reduction and interpretation of high resolution electrosprary mass spectra of large molecules," J. Am. Soc. Mass. Spectrom, 11, 320-332, 2000.

[5] Senko, M. W.; Beu, S. C.; McLafferty, F. W. "Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions," *J. Am. Soc. Mass Spectrom. 6*, 229-233, 1995.

[6] Senko, M. W.; Beu, S. C.; McLafferty, F. W. "Automated assignment of charge states from resolved isotopic peaks for multiply charged ions," *J. Am. Soc. Mass Spectrom. 6*, 52-56, 1995.

[7] A. Peter Snyder. "Interpreting protein mass spectra," Oxford University Press, 2000.