Bioinformatics and Biosystems 2006, Vol. 1, No. 2, pp. 109-114

질량스펙트럼의 펩타이드 분자량 오차범위 재해석에 의한 단백질 동정의 성능 향상

(Improvement of protein identification performance by reinterpreting the precursor ion mass tolerance of mass spectrum)

권 경 훈[1], 김 진 영[2], 박 건 욱[1], 이 정 화[2], 백 융 기[3], 유 종 신[1]

[1]한국기초과학지원연구원 연구장비개발부
[2]한국기초과학지원연구원 단백질체구조연구부
[3]연세프로테옴연구센터, 질병유전단백체연구지원센터

초 록

프로테오믹스에서 얻는 탄뎀 질량 스펙트럼은 효소로 가수분해된 펩타이드의 전구이온(precursor ion) 분자량과 펩타이드에 에너지를 가하여 생성된 이온조각(fragment ion)들의 분자량값들로 구성된다. 탄뎀 질량스펙트럼의 전구이온 분자량은 단백질 서열 데이터베이스에서의 검색 과정에서 가장 먼저 고려하는 값이다. 단백질 검색 프로그램은 단백질 서열 중에 스펙트럼의 전구이온으로부터 계산된 분자량과 일치하는 펩타이드 서열들을 찾아내고, 이들 중의 하나를 이온조각들의 분자량 정보를 이용해서 선택한다. 이 때에 전구이온의 분자량은 사용자가 지정한 오차범위 내에서 일치하는 값을 검색하는데, 이때의 오차범위는 질량분석기의 정확도에 따라 결정된다. 본 논문에서는 인간 혈액의 혈장시료로부터 FT LTQ 질량분석기를 통해 얻어진 탄뎀 질량 스펙트럼에서 전구이온 분자량의 분포를 역순서열을 이용하여 분석하였다. 전구이온 분자량의 분포를 재해석하여 실험값의 정확도를 보정하고 단백질 동정의 성능을 향상시키는 방법을 모색하였다.

키워드: 프로테오믹스, 탄뎀 질량 스펙트럼, 단백질 동정, FT LTQ 질량분석기, 역순 서열 데이터베이스

Abstract

In proteomics research, proteins are digested into peptides by an enzyme and in mass spectrometer, these peptides break into fragment ions to generate tandem mass spectra. The tandem mass spectral data obtained from the mass spectrometer consists of the molecular weights of the precursor ion and fragment ions. The precursor ion mass of tandem mass spectrum is the first value that is fetched to sort the candidate peptides in the database search. We look for the peptide sequences whose molecular weight matches with precursor ion mass of the mass spectrum. Then, we choose one peptide sequence that shows the best match with fragment ions information. The precursor ion mass of the tandem mass spectrum is compared with that of the digested peptides of protein database within the mass tolerance that is assigned by users according to the mass spectrometer accuracy. In this study, we used reversed sequence database method to analyze the molecular weight distribution of precursor ions of the tandem mass spectra obtained by the FT LTQ mass spectrometer for human plasma sample. By reinterpreting the precursor ion mass distribution, we could compute the experimental accuracy and we suggested a method to improve the protein identification performance.

Keywords: proteomics, tandem mass spectrum, protein identification, FT LTQ mass spectrometer, reversed sequence database

서 론

단백질 연구는 지난 10년간 프로테오믹스 분야의 발전에 의해 초고속으로 단백질의 통합정보를 생산할 수 있게 되었다. 2차원 전기영동법에 의해 단백질을 분리하고 질량분석기로 동정하는 분석 방법으로 시작된 프로테오믹스는 액체 크로마토그램과 질량분석 기술의 개발로 데이터의 정확도가 크

교신저자 : 유종신 (Email: jongshin@kbsi.re.kr)
본 연구는 보건복지부의 질병단백체연구센터(03-PJ10-PG6-GP01-0002 to YKP) 의 지원으로 수행되었음.

게 향상되었으며, 시료의 분석 속도도 나날이 발전하고 있다. 현재 초고속프로테오믹스 분석 방법으로 가장 일반적인 방법은 MudPIT (Multi-dimensional protein identification technology)이라 불리는 방법으로서, 가수분해된 펩타이드의 용액을 전하량, 소수성, 친수성 등의 펩타이드 특성에 따라 분리한 뒤에 질량분석기로 펩타이드를 분석함으로써, 양질의 데이터를 얻는다.(Gevaert 등, 2002) 특히 FT LTQ 질량분석기(ThermoFinnigan, San Jose, CA)는 이온 트랩 질량분석기의 고속 분석기능에 푸리에 변환 (Fourier Transform) 질량분석기의 정확도를 융합하여 정확한 데이터를 단시간에 얻을 수 있다.

MudPIT 방법에 의해 분리된 펩타이드 시료는 질량분석기에서 분자량이 측정되고, 다시 에너지를 받아서 조각이온(fragment ion) 들로 분해된다. 조각이온들의 분자량값들을 측정하면 탄뎀 질량 스펙트럼들이 얻어지는데, 한 실험에서 대략 수십만 개의 탄뎀 질량 스펙트럼을 얻는다. 탄뎀 질량 스펙트럼은 SEQUEST(Eng 등, 1994), Mascot (Perkins 등, 1999), X!Tandem (Craig, Beavis, 2004) 과 같은 데이터베이스 검색 프로그램에 의해 펩타이드 서열을 찾는 데에 사용된다. 그러나 탄뎀 질량 스펙트럼으로부터 펩타이드 서열을 동정하는 비율은 대개의 실험에서 10%~20% 에 불과하다. 이는 스펙트럼의 질적 문제와 펩타이드 서열의 동정 방법의 문제에 기인한다. 실험에서 얻은 스펙트럼으로부터 펩타이드 서열 정보를 최대한 얻어내기 위해서는 양질의 스펙트럼들에 대한 보다 다양한 분석이 필요하다. 이에 데이터베이스 검색 알고리듬과 데이터 분석 방법의 개선으로 펩타이드 서열 동정의 비율을 높이려는 노력이 계속되고 있다.(Nesvizhskii 등 2005)

본 논문에서는 고해상도로 양질의 데이터를 대량으로 얻는 FT LTQ 질량분석기에서의 펩타이드 분자량 측정 결과를 분석하고 장비의 특성을 활용하여, 탄뎀 질량 스펙트럼 데이터로부터 많은 단백질을 높은 신뢰도로 동정할 수 있는 방안을 모색하였다. FT LTQ 는 일반적으로 4ppm 의 해상도를 구현할 수 있는 질량분석기로 알려져 있으며 펩타이드의 분자량에 대해 매우 정밀한 결과를 제공한다. 이러한 고해상도를 십분 활용하여 단백질 동정의 효율을 높이기 위한 분석법을 제안하였다.

방법 및 알고리즘

프로테오믹스에서의 데이터베이스 검색 프로그램들은 단백질 서열 데이터베이스로부터 계산된 펩타이드들의 조각 이온 피크 값을 실험에서 얻은 스펙트럼 피크들과 비교하여 가장 잘 맞는 펩타이드를 골라낸다. 이를 위해서는 우선 단백질 데이터베이스에서 실험에서의 펩타이드 이온 분자량과 오차범위 내에서 일치하는 펩타이드 서열의 집합을 얻은 뒤에 이들의 조각 이온들의 분자량의 값들이 실험에서의 탄뎀 질량 스펙트럼과 가장 잘 일치하는 펩타이드를 선택한다. 여기서 펩타이드 이온의 분자량 비교는 스펙트럼을 비교할 대상이 되는 펩타이드들을 단백질 데이터베이스로부터 가려내는 과정이다. 데이터베이스 검색 프로그램에서 가장 많은 시간을 소요하는 부분은 스펙트럼과 펩타이드 서열을 비교하여 스코어를 계산하는 부분인데, 펩타이드 이온 분자량, 즉 전구이온 분자량의 오차범위를 확대하여 대상 펩타이드 수가 늘어나면, 데이터베이스 검색에 걸리는 시간은 그만큼 증가하게 되고, 실제 펩타이드와 무관한 펩타이드들이 비교대상으로 사용됨으로써 잘못된 펩타이드 서열을 선택할 가능성이 늘어난다. 반대로, 장비의 높은 정확도를 활용하여 오차 범위를 적절히 선택하여 이를 만족시키는 대상 펩타이드 수를 줄인다면, 데이터베이스 검색 속도가 현격하게 향상되는 효과가 있으며, 잘못된 펩타이드들이 선택될 확률이 줄어든다. 그러나, 오차범위를 너무 작게 택하면 올바른 펩타이드 서열이 대상에서 빠지게 되어 펩타이드 검색 효율이 낮아진다. 따라서 펩타이드 분자량의 오차범위의 적절한 설정은 데이터베이스 검색에서 실제 시료에 포함되어 있는 펩타이드의 데이터가 포함될 정도의 적정 범위를 결정하게 하고, 이에 따라 검색 효능과 정확도를 향상시키게 된다.

펩타이드 이온의 분자량 계산에는 펩타이드를 구성하는 각 아미노산들의 분자량을 사용하게 된다. 아미노산의 분자량은 자연계에 존재하는 수소, 탄소, 산소, 질소의 동위원소 존재비에 따라서 하나의 분자량 값이 아닌 여러 분자량의 값으로 존재한다. 해상도가 낮은 장비에서는 이와 같은 동위원소에 의한 분자량의 차이가 스펙트럼에서 구분되지 않으나, 고해상도의 장비에서는 각각의 분자량의 차이가 스펙트럼에서 드러나게 된다. 펩타이드 이온의 전하량과 동위원소 분포에 따른 스펙트럼 분석은 해상도가 높은 스펙트럼에서 monoisotope peak을 예측할 수 있게 한다.(Senko 등, 1995) 이러한 차이를 반영하여 일반적으로 데이터베이스 검색 프로그램은 탄뎀 질량 스펙트럼의 이온 피크들을 두 가지 경우로 분류하여 분석한다. 질량분석기가 동위원소의 분포를 구분할 수 있을 정도로 해상도가 높은 경우에는 동위원소 분포 중에 가장 낮은 분자량을 가지는 monoisotope peak 값을 데이터로 사용하고, 해상도가 낮은 경우에는 동위원소들의 분포가 구분되지 않으므로 동위원소 분포의 평균값을 검색 프로그램에서 사용한다. FT LTQ 질량분석기는 고해상도 질량분석기로서 그림 1 에서와 같이 동위원소들의 분포를 선명하게 확인할 수 있다. 고해상도 질량분석기의 경우에 동위원소에 의한 스펙트럼 피크의 분포를 살펴보면, 이온의 분자량이 작은 때에는 monoisotope peak 이 가장 큰 피크로 맨 앞에 나타나는데, 이온의 분자량이 커짐에 따라 monoisotope peak 의 상대적인 크기는 줄어들게 된다. 한편 대량프로테옴분석을 위해 data dependant MS/MS 모드에서 질량스펙트럼을 얻는 경우 일반적으로 이온의 세기가 가장 큰 이온이 전구이온으로 선택된다. 이온의 분자량이 작은 경우 전구이온이 monoisotope 이온과 일치하여 정확한 분자량의 펩타이드가 검색되지만 이온의 분자량이 증가함에 따라 monoisotope 다음의 이온의 세기가

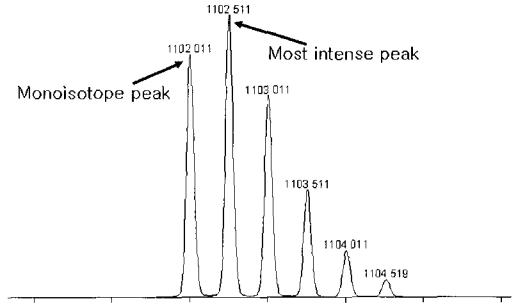

그림 1. Monoisotope peak 와 most intense peak 의 비교 예. 그림의 스펙트럼은 펩타이드 서열 SDLEBQYETLQEELBALK 의 전구이온 피크들로, 분자량은 2202.0071 Da 이고 전하량은 2+ 이다. 여기서 첫 번째 피크인 monoisotope peak 이 두 번째 피크보다 작으므로, 질량분석기에서 자동으로 전구이온을 선택할 때에는 두 번째 피크를 선택하기도 한다.

더 커지고 전구이온으로 선택되게 된다. 이러한 경우에는 전구이온에 의해 계산된 펩타이드 이온 분자량이 실제 펩타이드 분자량과 +1 Da 또는 +2 Da 만큼 차이가 나게 된다.

펩타이드 이온 분자량의 오차는 monoisotope 이온이 전구이온으로 선택되지 못함으로써 분자량을 잘못 판정하게 되어 +1 Da, +2 Da 주위의 값들이 나타나는 것으로, 데이터베이스 검색 스코어의 분포를 살펴보면 검색 스코어가 높은 펩타이드들이 0 Da 뿐만 아니라, +1 Da, +2 Da 에 상당량 분포함을 볼 수 있다. FT LTQ 질량분석기의 전구이온 분자량 오차범위에 대한 분석에서는 한국인의 혈장을 시료로 사용하였고, 데이터베이스 검색은 TurboSEQUEST (ThermoFinnigan, San Jose, CA) 프로그램을 사용하였다. TurboSEQUEST에서는 cross correlation value (Xcorr)라는 검색 스코어를 얻는데 이 스코어만으로 검색 결과를 판단하기는 어려우므로, 그 외에 DCn, RSp 등의 스코어 값을 보조적으로 활용한다. 본 연구에서는 TurboSEQUEST에서 질량분석 스펙트럼과 관련된 스코어들을 통합하여 정의한 스코어로서 Keller 등이 정의한 F 스코어 (Keller 등, 2002)를 사용하였다. F 스코어가 작은 값인 경우에는 펩타이드 서열을 동정하기는 하였으나 스펙트럼과 펩타이드 서열이 그다지 잘 맞지 않는 경우들로서 false positive 데이터들을 많이 포함한다. F 스코어가 높은 펩타이드는 정확한 검색이 이루어졌음을 나타내는데, FT LTQ 데이터의 검색 결과를 분석한 결과 높은 검색 스코어값들이 분자량의 오차가 0 Da, 1 Da, 2 Da 인 영역 주변에 특히 많이 존재함을 볼 수 있다. F 스코어 값이 4보다 크게 검색된 펩타이드의 99%가 0 Da, 1 Da, 2 Da 지점으로부터 0.03Da 내의 영역에 존재하였다. 이는 FT LTQ 질량분석기에서 펩타이드 이온의 두 번째 피크, 또는 세 번째 피크를 전구이온으로 선택하여 전구이온으로부터 계산된 펩타이드 분자량과 검색된 펩타이드의 분자량과의 오차가 1Da, 2Da 으로 커지는 경우가 발생함을 보여준다. 많은 실험실들에서 FT LTQ 가 고해상도의 장비이므로 탄뎀 질량 스펙트럼의 데이터베이스 검색에서 1 Da 보다 작은 오차범위 값으로 펩타이드를 검색하는데, 그러한 경우에는 1 Da, 2 Da 영역 주변의 펩타이드들의 검색 결과를 버리게 된다.

우리는 이러한 monoisotope peak 의 오차가 전구이온의 분자량이 큰 경우에만 국한된 현상인지 아닌지를 확인하기 위해 펩타이드 이온의 분자량에 따른 분자량 오차값의 분포를 살펴보았으나, 오차값이 크면서 높은 F 스코어를 얻은 펩타이드들이 큰 분자량에만 존재하는 것은 아니며, 1000 Da ~ 3000 Da 의 펩타이드 분자량 영역에서 모두 나타나는 현상이었다. 우리는 전구이온 분자량의 오차범위와 관련한 이같은 사전 분석 결과들을 토대로 FT LTQ 질량분석기의 데이터로부터 보다 정확하고 효율적인 단백질 동정 결과를 얻기 위한 분석 시스템을 구성하였다.

시스템 구조

1. 역순 서열(reversed sequence) 데이터베이스 활용

탄뎀 질량 스펙트럼의 데이터베이스 검색은 TurboSEQUEST 프로그램을 사용하고, 데이터베이스로는EBI (European Bioinformatics Institute, http://www.ebi.ac.uk/) 에서 제공하는 IPI human database를 사용하였다. 검색 결과 얻은 펩타이드의 진위를 파악하기 위하여 IPI 데이터베이스의 역순 서열 데이터베이스를 함께 활용하였다.

역순 서열 데이터베이스는 단백질의 아미노산 서열을 뒤집어서 C-터미날의 아미노산부터 N-터미날의 아미노산까지로 서열을 만든 가상의 단백질 데이터베이스이다. 단백질 데이터베이스와 역순서열 데이터베이스를 합한 데이터베이스로 탄뎀질량 스펙트럼을 검색하면 검색 결과가 단백질 데이터베이스에서 얻어진 펩타이드와 역순서열에서 얻어진 펩타이드들이 얻어진다. 이들의 분포를 그려보면, 검색 스코어가 낮은 경우에는 역순서열에서 얻은 펩타이드의 스코어 분포와 단백질 서열에서 얻은 펩타이드의 스코어 분포가 같다. 검색 스코어가 낮은 분포는 잘못된 펩타이드 서열임을 뜻하는데, 실제로는 역순 서열에서 동정된 펩타이드의 개수만큼 단백질 서열에서 false positive 결과를 얻음을 알 수 있다. 이러한 성질은 역순 서열의 데이터베이스가 단백질 서열과 크기가 같고 펩타이드 분자량의 분포도 같음으로써 일어난다. 역순 서열은 무작위 서열 데이터베이스와 같이 실제 단백질 서열과는 다른 펩타이드 서열을 보여줌과 동시에 펩타이드 분포가 실제 단백질 데이터베이스와 동일한 데이터베이스이므로, 질량분석 데이터의 분석에 매우 유용하다.

펩타이드 검색에서 역순 서열 데이터베이스를 이용한 false positive 비율의 계산은 최근에 프로테오믹스 실험실들에서 많이 시도되고 있다.(Elias 등, 2005; Park 등, 2006) 이러한 검색 방법은 역순 서열 데이터베이스를 미끼로 단백질 데이터베이스에서의 false positive 를 가려내는 방법이라 하여 'decoy' 방법이라 부른다. 이에 대한 변칙적인 방법으로는

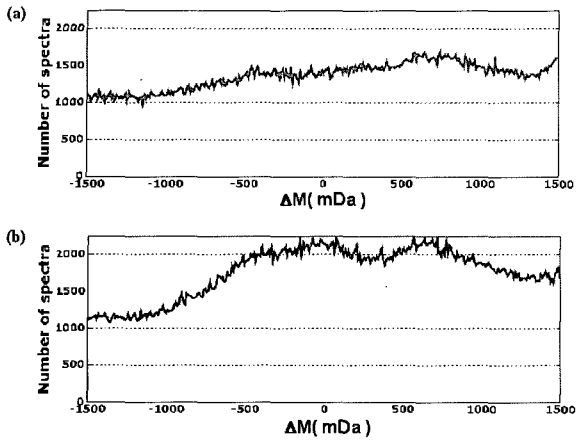

그림 2. LT LTQ/MS 에서 전구이온의 분자량과 검색된 펩타이드 서열의 분자량의 차이값의 분포 (a) 역전서열 단백질 데이터베이스에서 검색된 펩타이드의 분자량 차이 분포 (b) 단백질 데이터베이스에서 검색된 펩타이드의 분자량 차이 분포

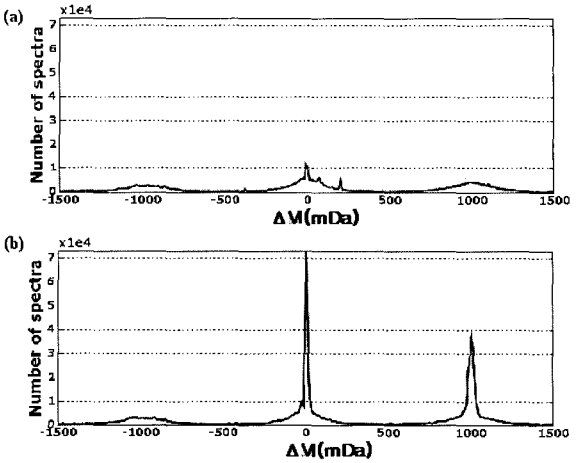

그림 3. FT LTQ/MS 에서 전구이온의 분자량과 검색된 펩타이드 서열의 분자량의 차이값의 분포 (a) 역전 서열 단백질 데이터베이스에서 검색된 펩타이드의 분자량 차이 분포 (b) 단백질 데이터베이스에서 검색된 펩타이드의 분자량 차이 분포

'shuffle' 방법이라 하여, 역순 서열 데이터베이스의 서열에서 펩타이드의 아미노산 순서를 바꿈(shuffle)으로써 서열의 임의성을 강화하는 방법이 있다. 하지만 shuffle 방법에 의한 분석 결과와 원래의 decoy 방법에 대한 분석 결과는 크게 차이가 없다고 보고되고 있다.

그림 2와 그림 3 은 각각 LT LTQ 와 FT LTQ에서 검색한 펩타이드의 분자량 오차값의 분포를 보여준다. 두 그림에서 (a) 의 분포는 각각의 질량분석기에서 역순 서열에서 얻어진 펩타이드의 분포들이고, (b) 의 분포는 실제 단백질 데이터베이스로부터 얻어진 펩타이드들의 분포이다. 여기서 LT LTQ 는 펩타이드의 분자량과 탄뎀 질량 스펙트럼을 모두 이온트랩 장치로 얻는 질량분석 방법을 뜻하며, FT LTQ 는 펩타이드 분자량을 FT 방법으로 측정하고 탄뎀 질량 스펙트럼은 이온 트랩 장치로 얻는 방법이다. 이 때에 FT LTQ 의 경우는 고해상도로 전구이온의 분자량을 측정할 수 있다. 때문에 그림 2에서는 (a), (b) 두 곡선의 관계가 명확하게 드러나지 않지만, 그림 3에서는 (b) 의 곡선이 (a) 의 분포에서 0 Da, 1 Da 주변에만 펩타이드들이 추가된 모양임을 알 수 있다. 이처럼 역순 서열 분포와 단백질 서열 분포가 일부 영역에서는 확연히 다르고 나머지 영역에서는 동일한 경우에 우리는 차이가 나는 영역에 true positive 들이 모여있음을 예측할 수 있다. 즉 역순 서열은 시료에 포함되지 않은 임의의 펩타이드이므로, 단백질 데이터베이스로부터 올바르게 동정된 펩타이드들이 그림 3(b)에서 0 Da, 1 Da 에 모여 있음을 나타낸다.

역순 서열 데이터베이스에서 동정된 펩타이드 서열의 분포는 단백질 데이터베이스에서의 틀린 서열, 즉 false assignments 의 분포와 같다. 다시 말하면, 단백질 데이터베이스에서 검색된 펩타이드 중에 옳은 펩타이드의 분포는 즉 그림 2 또는 3의 (b) 곡선에서 역순 서열 데이터베이스의 분포, 즉 그림의 (a) 곡선을 뺀 분포로 표현할 수 있다. 그러므로, 우리는 그림 3의 분포도를 사용하여 펩타이드 검색 결과 중에 몇 퍼센트가 옳은 결과인지를 판단할 수 있다. 그림 4는 역순 서열에서 얻은 펩타이드의 검색 스코어 분포를 틀린 펩타이드의 점수 분포로 그리고, 올바른 펩타이드의 점수 분포는 단백질 데이터베이스에서 얻은 펩타이드의 검색 스코어 분포곡선에서 역순 서열 데이터베이스의 펩타이드 분포를 뺀 곡선으로 계산하여 얻은 그래프이다. 이 그림에서 오른쪽의 올바른 펩타이드 분포곡선과 왼쪽의 틀린 펩타이드 곡선이 겹치는 부분은 검색 스코어만으로는 검색 결과의 옳고 그름을 판단할 수 없는 영역이다. 이 영역에 대해 우리는 오류율을 계산하고, 검색된 펩타이드가 오류율 이내에서 맞는 결과라고 결론짓는다. 그림 4 에서는 오류율 5% 내의 최소 검색 스코어를 설정한 예를 보여준다.

참고로, 그림 3(a)에서 역순 서열의 펩타이드 분자량의 오차가 정수 값의 주위에 모여 있는 것은 펩타이드가 탄소, 산소, 수소, 질소들의 화합물로 얻어지는데, 이는 각 원소의 원자량 값이 거의 정수에 가까운 값이어서 나타나는 현상이다. 실제로 질량 스펙트럼과 무관하게, 일반 단백질 서열 데이터베이스에서 트립신으로 가수분해된 펩타이드의 분자량 값의 분포를 그려보면, 모든 분자량 영역에 골고루 분포하지 않고 약 1 Da 의 간격으로 떨어져서 분포한다. 이는 질량분석 실험에서 측정하는 펩타이드가 7~30 개 정도의 아미노산으로 구성되는데, 각 아미노산의 분자량은 모두 정수값에서 0.03 ~ 0.1 Da 정도 더해진 값을 가지므로, 이들 아미노산 분자량의 합인 펩타이드의 분자량 값도 아미노산 서열에 따라서 정수값 정도의 차이를 보이게 된다. 해상도가 낮은 LT LTQ 데이

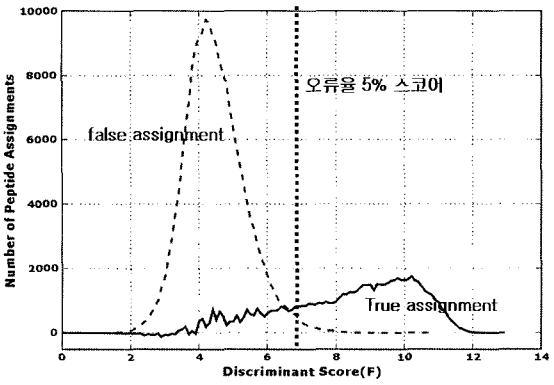

그림 4. 역순 서열 데이터베이스에서 얻은 펩타이드 분포를 통하여 간접적으로 추정된 맞는 펩타이드와 틀린 펩타이드의 분포. 점선은 역순 서열 데이터베이스의 분포이면서 false assignment 의 분포이다. 그림과 같은 양상의 분포는 LCQ 나 LTQ에서 대개 나타나는 분포이다. 두 펩타이드 분포 곡선으로부터 5% 오류범위 이내의 펩타이드 영역을 결정할 수 있다. FT LTQ 의 경우에는 전구이온 분자량의 오차 범위에 따라서 false assignments 의 개수가 크게 변화하여, 분자량 오차범위가 작은 경우에 true assignments 보다 false assignments 가 훨씬 작은 분포를 보일 수 있다.

터의 경우인 그림 2(a) 에서는 장비의 해상도가 낮아서 전구이온의 분자량이 정확히 측정되지 않아 이러한 경향이 거의 눈에 띄지 않았다.

본 연구에서는 인간 혈장 시료에 대해서 이와 같이 IPI 단백질 데이터베이스와 이의 역순 서열 데이터베이스을 활용하여 FT LTQ 질량스펙트럼에서의 true positive 펩타이드의 분포를 얻을 수 있었다. 지금까지 발표되어진 다른 분석 방법들에서는 역순 서열 데이터베이스의 검색 결과를 검색 스코어에 적용하여 펩타이드 동정의 오류율을 계산하는 데에만 사용하였으나, 본 연구에서는 전구이온 분자량의 오차범위에 따른 펩타이드의 분포 분석에 역순 서열 데이터베이스의 검색 결과를 활용하여 탄뎀 질량 스펙트럼의 오차범위를 명백하게 가시화하여 설정할 수 있었다. 이러한 분석 방법은 LCQ, LT LTQ 와 같이 해상도가 낮은 장비에서는 역순 서열에서 얻은 펩타이드 분포가 일정한 모양을 가지지 못하므로, 이러한 분석 방법은 FT LTQ에서 특히 유용한 방법이다.

2. 전구이온 분자량의 오차범위 설정

역순 서열을 통해 걸러낸 true positive 펩타이드들로부터 적절한 전구이온 오차범위 값을 정하기 위하여는 0 Da, 1 Da, 2 Da 의 주변에 분포한 펩타이드들의 분포를 알아보아야 한다. 우리는 펩타이드 분자량의 오차값이 1 Da 에 가까운 분포는 1 Da 만큼 분포곡선을 왼쪽으로 이동하고, 2 Da 주위의 분포는 2Da 만큼 왼쪽으로 이동하여 0 Da 주위에 분포곡선을 모으게 하였다. 만일 FT LTQ 질량분석기에서 monoisotope peak가 정확하게 측정이 된다면 오차값의 분포는 0 Da 근처에 모이게 될 것이다. 우리는 이렇게 모은 펩타이드의 분포로부터 true positive 분포 곡선을 얻고, 이를 다시 가우스 함수분포로 보정하여 전체 펩타이드 중에 95% 펩타이드를 포함하는 오차범위 영역을 계산하였다. 이때 전구이온 분자량의 오차 범위는 -17 mDa ~ 26 mDa 으로 얻어졌다.

monoisope peak 에 의해 동정된 탄뎀 질량 스펙트럼은 34,038 개이고, M+1, M+2 이온 피크에 대하여 33,530 개의 스펙트럼이 추가로 동정되었다. 이들 중에 서로 다른 펩타이드의 서열은 monoisotope peak 의 경우에 2,663개, M+1, M+2 이온 피크를 포함하여 검색된 펩타이드는 4,551 개였으며, 이로부터 동정한 단백질의 개수는 monoisotope 이온에서 269 개이던 것이 M+1, M+2 이온을 추가함으로써 414 개로 증가하였다. 우리는 역순서열 데이터베이스을 이용한 false positive 의 제거, M+1이온과 M+2 이온을 고려한 오차범위 분석에 의해 145 개, 즉 54% 의 단백질을 95%의 신뢰도로 추가로 동정하였다.

결 론

종래에 발표되었던 펩타이드 서열 검색에 대한 신뢰도의 분석은 대부분 LCQ, LT LTQ 와 같이, 대량의 데이터를 생산하지만 해상도가 낮은 장비에서 얻은 데이터에 대한 분석이었다. FT LTQ 는 이들 장비와 달리 높은 해상도를 보이므로, 전구이온 분자량의 오차범위를 매우 작게 설정할 수 있다. 전구이온 분자량의 오차범위를 축소하면, 단백질 서열 데이터베이스에서 이 오차범위를 만족시키는 임의의 펩타이드가 존재할 확률이 매우 작아지므로, 데이터베이스 검색 결과 얻은 스코어는 LCQ, LT LTQ 에서와는 다른 검색 스코어 분포를 보인다. 이에 따라 false assignment 의 분포가 상대적으로 작아질 수 있다.

본 논문은 LCQ, LT LTQ 질량분석기의 데이터에 대해 개발되어온 분석법을 활용하여 FT LTQ 의 질량 스펙트럼을 분석하고, FT LTQ 의 높은 정확도를 적용하여 보다 효율적인 분석 결과를 유도할 수 있는 방법을 알아보았다.

우리는 역순 서열 데이터베이스의 분포를 전구이온 분자량의 오차분포에 적용함으로써 데이터베이스 검색 결과의 진위 분석을 오차분포별로 계산할 수 있었다. 작은 오차범위에서 FT LTQ 의 데이터에서는 옳은 펩타이드 동정의 개수에 비해서 틀린 펩타이드 동정의 개수가 더 적었음에도 불구하고, 단백질 데이터베이스에서의 펩타이드 분포와 역순 서열 데이터베이스에서의 펩타이드 분포 양상으로부터 기존의 통계적인 해석 방법을 그대로 적용하여 false positive 의 개수를 예측할 수 있었다. 뿐만 아니라, monoisotope peak 와 더불어 두 번째, 세 번째 피크 값이 선택되는 경우를 전구이온 분자량의 오차범위에 포함함으로써 54% 만큼 더 많은 단백질을 동정

할 수 있었다.

이러한 단백질 동정 성능의 차이는 전구이온과 monoisotope 이온이 일치하지 않은 경우 발생하는 문제이다. 여기서는 문제의 해결 방법으로 오차범위의 조정을 제시하였다. 이에 대한 또다른 해결 방법으로는 monoisotope peak 을 설정하는 알고리듬의 개선이 가능하다. 동위원소 분포로부터 전구이온의 스펙트럼을 정확히 예측함으로써, 가장 큰 peak 을 선택하면서도 monoisotope peak 은 원래의 정확한 값을 계산한다면 펩타이드 서열 동정을 보다 효율적으로 처리할 수 있다. 이러한 작업은 질량분석기의 운영 시스템에서 수정되어야 할 부분이다. 데이터 분석 차원에서 다른 해결 방법으로는 스펙트럼 데이터의 보정을 생각할 수 있다. 스펙트럼 데이터에서 조각 이온의 스펙트럼은 바꾸지 않은 채로 전구이온의 분자량 값에 1 Da 또는 2 Da 을 더한 값들로 이루어진 질량 스펙트럼을 추가하는 방법이다. 이러한 방법은 본 논문에서 사용한 방법과 거의 유사한 결과를 보이리라고 예상된다.

질량분석 기술의 발전과 함께 장비의 정확도 및 측정 속도는 계속 향상하고 있다. 이에 따라 데이터의 분석 방법도 계속적인 진화가 필요하다. 낮은 해상도에서 적용되던 분석 방법들이 고해상도에서는 더 이상 적용되지 않는 경우가 생길 것이며, 낮은 해상도에서는 불가능했던 분석이 장비의 개선으로 가능해지는 경우도 있다. 본 연구에서와 같이 기존의 낮은 해상도에서 사용하던 분석 방법을 재고하고 재해석하는 과정은 신기술, 첨단 장비를 위한 새로운 분석 방법을 고안하는 계기가 될 수 있을 것이다.

참 고 문 헌

[1] Gevaert, K. et al. (2002) 'Chromatographic isolation of methionine-containing peptides for gel-free proteome analysis: identification of more than 800 Escherichia coli proteins', *Mol. Cell. Proteomics,* 1: 896-903.

[2] Nesvizhskii, A.I., et al. (2006) 'Dynamic Spectrum Quality Assessment and Iterative Computational Analysis of Shotgun Proteomic Data: Toward More Efficient Identification of Post-translational Modifications, Sequence Polymorphi는, and Novel Peptides', *Mol. Cell. Proteomics,* 5: 652-670.

[3] Senko, M.W., Beu, S.C., McLafferty, F.W.(1995) 'Automated Assignment of Charge Stastes from Resolved Isotopic Peaks for Multiply Charged Ions', *J. Am. Soc. Mass Spectrom.* 6: 52-56.

[4] Eng J.K., McCormack A.L., Yates J.R. III (1994) 'An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database', *J. Am. Soc. Mass Spectrom.* 5: 976 - 989.

[5] Perkins D.N., Pappin D. J., Creasy D. M., Cottrell J. S. (1999) 'Probability-based protein identification by searching sequence databases using mass spectrometry data', *Electrophoresis*, 20: 3551-3567.

[6] Craig, R., Beavis, R.C. (2004) 'TANDEM: matching proteins with tandem mass spectra', Bioinformatics, 20: 1466-1467.

[7] Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) 'Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search'. *Anal. Chem.*74, 5383 - 5392.

[8] Elias, J.E., *et al.* (2005) 'Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations', Nature Methods, 2: 667-675.

[9] Park, G.W. et al. (2006) 'Human plasma proteome analysis by reversed sequence database search and molecular weight correlation based on a bacterial proteme analysis', *Proteomics* 6: 1121-1132.