프로테오믹스 연구를 위한 정량분석 데이터의 해석

Data analysis for quantitative proteomics research

권 경 훈

한국기초과학지원연구원 연구장비개발부

Kyung-Hoon Kwon, Ph.D

Division of Instrument Development, Korea Basic Science Institute

52 Yeoeun-dong, Yusung-ku, Daejeon 305-333, Korea

khoon@kbsi.re.kr

초 록

프로테오믹스는 생물체 안에 포함되어 있는 단백질을 통합적으로 연구한다. 단백질을 동정(Protein identification)하고, 단백질의 상태를 분석(Protein characterization)하며, 단백질의 양적 변화를 관찰(Protein quantitation)한다. 단백질에 대한 분석, 특히 질량분석기에 의해 초고속으로 대량의 단백질 데이터를 생산하는 프로테오믹스의 연구는 정량적인 단백질 발현양상 분석의 정확도를 높이고 분석시간을 단축하기 위해 다양한 실험기법과 데이터 분석기법을 동원하고 있다.[1] 단백질의 양적 차이나 양적 변화의 관찰은 바이오마커를 발굴하고 생명현상의 메카니즘을 규명하여 그 결과를 신약개발에 활용하기 위한 기초 연구이다.

이 글에서는 프로테오믹스 연구의 초창기부터 사용되어온 2차원 전기영동법에 의해 생성되는 2D-gel image 에서의 스팟(spot) 분석법과 함께, 탄뎀 질량분석기를 사용하는 ICAT, SILAC 등의 동위원소를 사용한 라벨링(labeling) 방법, 라벨링을 하지 않는 label-free 방법 등 프로테오믹스에서의 정량분석법에 대한 기본 개념을 살펴보고, 이들에서의 데이터 분석 기술의 적용에 대해 간략히 소개하였다.

Keyword: 초고속 프로테오믹스, 단백질의 정량분석, 탄뎀 질량분석기, 바이오마커 발굴, 2D-gel

I. 서 론

전기영동법에 의해 단백질을 2차원으로 분리하여 2D-gel 을 얻은 뒤에 분리된 각각의 단백질 스팟을 질량분석기로 분석하고 단백질의 정보를 밝혀내는 연구는, 단백질의 정보를 한꺼번에 얻는다는 특징을 반영하여 기존의 프로테인 연구와 구분짓는 '프로테오믹스 (Proteomics)' 라는 이름을 얻었다. Omics 의 한 분야로, 대량의 데이터를 단시간에 생산하는 프로테오믹스는 첨단 실험장치와 분석기법의 개발로 발전을 거듭하여 초고속 프로테오믹스(High-throughput proteomics) 기술로 진화하였다. 초기의 프로테오믹스는 생물 시료에 어떤 단백질들이 존재하는지, 단백질을 정확하게 동정하는 데에 초점이 맞추어졌으며, 당시에는 수백 개의 단백질을 한꺼번에 동정해낸다는 사실만으로도 생물학자들에게는 충분히 매혹적인 분야가 되었다.

하지만, 프로테오믹스의 궁극적인 목표는 단백질의 정량분석에 있다. 단백질들이 세포 내의 환경 변

1) Hotchkiss, J.L., Simpson, R.J., 'Proteins and Proteomics: A Laboratory Manual', Cold Spring Harbor Laboratory Press, 2003. ; 주현, 한진, '프로테오믹스: 단백질에 대한 이해 및 기능 해석의 새로운 접근과 응용', 범문사, 2004.

화에 따라 어떤 변화를 보이는지, 시료에 따라 어떤 차이를 보이는지를 단백질 하나하나에 대한 분석이 아닌, 단백질 전체, 즉 프로테옴(proteome) 의 패턴 변화에 의해 분석하고자 한다. 프로테오믹스에서의 정량분석은 2D-gel에서 스팟의 크기와 진하기로 단백질의 양을 가늠하는 기본적인 분석 방법과 펩타이드의 질량스펙트럼으로부터 단백질의 양적 정보를 계산하는 방법으로 분류하여 생각할 수 있다. 각각의 방법을 살펴보기로 하자.

II. 2D-gel 에 의한 정량분석

2D-gel 에서의 스팟들은 전기영동법에 의해 분리된 단백질이며, 서로 다른 상태에서의 2D-gel 스팟 분포를 비교하면 두 상태에서의 단백질을 비교할 수 있다. 두 개의 2D-gel을 겹쳐서 특정 단백질의 스팟을 비교하기 위해서는 두 gel에서 동일한 스팟들을 짝짓기 위한 정렬(alignment) 작업과 gel 에 전체적으로 분포해 있는 백그라운드를 제거하는 작업, 그리고, 스팟의 강도에 대한 정규화(normalization) 작업이 필요한데, 이런 일들은 대부분 2D-gel image 분석 소프트웨어에서 일반 영상분석 알고리듬을 사용하여 제공한다. 이렇게 2D-gel 의 각 스팟에 대한 강도가 분석 소프트웨어에서 계산되고 나면, 마치 마이크로어레이의 각 스팟에서 유전자의 발현량 변화를 비교하듯이 스팟의 강도 변화를 비교하여 변화 패턴이 유사한 단백질들을 그룹으로 분류할 수 있다. 단백질 스팟의 분류 방법으로는 Principal Component Analysis, Singular Vector Machine, Singular Value Decomposition 등의 알고리듬이 사용된다.[2] 그러나, 2D-gel 스팟들을 강도의 변화 패턴에 따라 분류하는 알고리듬이 의미 있는 결과를 주기 위해서는 대상 스팟의 개수가 충분히 많아야 하며, 스팟의 강도가 재현성 있게 나타나야 한다. 최근들어 전기영동법의 재현성이 높아지고 감도가 향상됨에 따라, 2D-gel 결과로부터 단백질의 발현 프로파일을 분석하여 다른 단백질과는 구분되는 패턴으로 변화를 보이는 단백질 그룹을 찾아내고 이를 질병관련 바이오마커와 연관 짓는 연구들이 발표되고 있다.[3] 특히 DIGE(Difference Gel Electrophoresis) 방법은 2D-gel에서 두 시료에 서로 다른 형광염료를 반응시킨 후에 시료를 섞어서 전기영동으로 2D-gel을 얻는 방법으로 2D-gel 의 감도와 재현성의 문제 해결에 많은 기여를 하였다.[4]

III. 질량분석기에 의한 정량분석

2D-gel에서 분리된 단백질의 양을 비교하는 방법과는 달리, 질량분석에서는 효소에 의해 분해된 펩타이드의 양에서 간접적으로 단백질을 비교한다. 하나의 단백질에 속한 여러 개의 펩타이드에서 제각기 얻은 정량분석 값들은 동일한 단백질로부터 생성되었음에도 불구하고 실험오차에 의해 약간씩의 차이를 보이게 된다. 펩타이드간의 양적 차이는 분석결과의 신뢰도를 판단하는 기초자료로 활용될 수 있다. 그렇지만, 몇몇 단백질에 공통으로 존재하는 펩타이드 또는 일부가 수식화에 의해 변형된 펩타이드의 양은 해당 단백질의 전체량을 반영하지 않으므로, 별도의 해석이 요구된다.

1. Protein Abundance Index

화학적으로 라벨링 방법에 의해 시료에 태그를 붙

2) Kuruvilla, F.G., Park, P.J., Schreiber, S.L., 'Vector algebra in the analysis of genome-wide expression data', *Genome Biol.* 2003, 2(3), research 0011.1-0011.11.

3) Gerlling, I.C., Singh, S., Lenchik, N.I., Marshall, D.R., 쩌, J., 'New Data-Analysis and Mining Approaches Identify Unique Proteome and Transcriptome Markers of Susceptibility to Autoimmune Diabetes', Mol. Cell. Proteomics 2006, 5: 293-305. : Stein, R.C., Zvelevil, M.J., 'The application of 2D gel-based proteomics methods to the study of breast cancer', *J. Mammary Gland. Biol. Neoplasia*, 2002, 7: 385-93.

4) Marouga, R., David, S., Hawkins, E., 'The development of the DIGE system: 2D fluorescence difference gell analysis technoogy', *Anal. Bioanal. Chem.* 2005, 382:669-78.

이는 정량분석 방법이 보편화되기 전에 질량분석 데이터로 단백질을 정량하는 데에는 Protein Abundance Index(PAI) 방법이 사용되었다.[5] PAI 는 시료에 많이 포함되어 있는 단백질에 포함되어 있는 펩타이드 서열이 탄뎀 질량스펙트럼에서 많이 발견된다는 가정에 근거한다. 이에 따라 PAI는 단백질에 포함된 펩타이드가 질량 스펙트럼에서 발견된 횟수로 단백질의 양을 추정한다. 이때 단백질에서 가능한 펩타이드가 많은 경우에는 펩타이드를 발견할 확률이 크므로, 발견 가능한 펩타이드 숫자로 질량스펙트럼의 개수를 나눈 값을 PAI로 정의한다.

PAI 는 정량분석을 위한 별도의 실험 없이 간단한 계산으로 단백질의 양을 비교할 수 있는 방법이지만, 오차범위가 커서 단백질 발현량의 적은 차이를 구분하는 목적으로는 적합하지 않다. PAI를 기초로 한 다른 측정지수도 고안되고 있으나, 프로테오믹스 실험에 관련된 여러 가지 복잡한 요소들을 고려하여 보다 면밀한 분석과 연구를 거쳐서 정량을 위한 지수를 개발해야 보다 유용하게 사용될 것이다.

2. 동위원소를 이용한 정량분석 방법

질량분석기로 프로테옴을 정량적으로 분석하는 방법은 주로 동위원소를 이용한다. ICAT™(Isotope-Coded Affinity Tagging) 방법이 그중 대표적인 예이다.[6] ICAT은 그림 1 에서와 같이 비교하고자 하는 두 시료에 수소를 함유한 시약과 중수소를 함유한 시약을 각각 처리하여 시스테인에 두 종류의 태그가 붙도록 한다. 두 시료에서의 동일한 아미노산 서열을 가진 펩타이드가 두 태그에 의해 분자량에 차이가 생기면, 질량분석기에서 이를 구분하여 두 시료에서 각 단백질의 발현 정도를 비교할 수 있다. 한편 최근들어 각광을 받는 정량분석 방법인 SILAC(Stable Isotope Labeling by Amino Acids in Cell Culture)[7] 은 ICAT 에서와 비슷하게 동위원소를 이용한 단백질의 정량분석 방법이나, ICAT과의 중요한 차이점은 세포 배양 전 단계에서 탄소의 동위원소로 처리한다는 점이다. 동위원소를 이용한 정량분석 데이터의 해석은 질량 스펙트럼에서 일정한 간격만큼 떨어진 동일 펩타이드들을 골라내어 피크의 크기를 측정함으로써 단백질의 발현량을 계산한다.

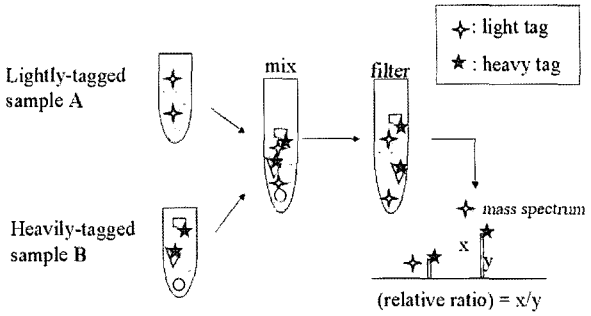


그림 1. 동위원소가 치환된 태그를 이용한 정량분석 방법. 두 시료에 수소(^{1}H)와 중수소(^{2}H) 또는 탄소(^{12}C) 와 탄소 동위원소(^{13}C)를 함유한 태그를 펩타이드의 특정 아미노산에 부착시킨 뒤에 두 시료를 합하고 태그된 펩타이드만 걸러내어 질량분석기로 스펙트럼을 얻으면, 태그의 분자량 차이를 이용하여 단백질의 양을 계산할 수 있다.

이처럼 동위원소로 펩타이드 분자량에 차 등을 둔 뒤에 두 시료를 섞어서 질량분석기로 질량스펙트럼을 얻는 방법은 두 차례의 반복실험에 의해 발생하는 오차를 줄이려는 의도로 고안된 방법이다. 실제로 동일한 시료를 가지고 같은 실험을 반복한다 해도 실험 오차에 의해 결과는 차이가 나므로, 오차 발생 요인을 줄이기 위하여 두 시료를 섞어서 한 번에 실험하는 프로토콜을 사용한다. 하지만, 이러한 방법에서도 오차 요인은 항상 존재하며, 이를 보정하는 정규화 작업이 필요하다. 이때의 정규화 작업은 두 시료에서 상당 부분 많은 단백질들의 발현량이 변하지 않는다는 가정하에 이루어진다. 즉, 두 시료가 크게 다르지

5) Rappsilber, J., Ryder, U., Larmond, A.I., Mann, M., 'Large-scale proteomic analysis of the human spliceosome', *Genome Res.* 2002, 12: 1231-1245.

6) Gygi, S.P. Rist, B., Griffin, T.J., Eng, J., Aebersold, R., 'Proteome analysis of low-abundance proteins using multidimensional chromatography and isotope-coded affinity tags', *J. Proteome Res.* 2002, 1: 47-54.

7) Ong, S.E., Kratchmarova, I., Mann, M., 'Properties of 13C-substituted arginine in stable isotope labeling by amino acids in cell culture (SILAC)', *J. Proteome Res.* 2003, 2: 173-81.

는 않은 상황으로 대다수의 단백질에는 발현량의 차이가 없다는 가정을 이용하면, 전체 시료에서 나타나는 공통적인 오차요인을 정규화 작업으로 보정할 수 있다. 두 시료에 대한 정량 데이터로부터 공통적인 오차요인을 제거하여 정규화 하는 방법은 동정된 펩타이드들에 대한 scatter plot을 사용한다.

x축을 시료 A 에서의 펩타이드 양, y축을 시료 B 에서의 펩타이드 양으로 하여 그래프에 한 점으로 펩타이드를 표시하면, 분석된 펩타이드들의 scatter plot을 얻으며, 대개의 점이 하나의 직선 주변에 분포함을 볼 수 있다. 이 직선의 기울기를 1로 보정하여 y=x 위의 직선 위에 점들이 놓이도록 데이터를 보정하면, 그 직선에서 크게 벗어난 펩타이드들이 시료 A와 B 에서 발현량에 차이를 보이는 펩타이드이다.

3. 태그를 사용하지 않는 정량분석 방법

최근 들어 질량분석기의 정확도와 재현성이 향상됨에 따라 태그를 사용하지 않고 두 시료를 독립된 두 번의 실험에서 분석하여 데이터를 비교하는 정량분석 방법이 가능해지고 있다.[8] 대량의 스펙트럼 데이터를 정렬하고 정규화해야 하므로 분석 시스템이 복잡하고 계산시간이 많이 걸리지만, 태그를 붙이기 위한 화학적인 처리 과정을 생략함으로써 오차를 줄

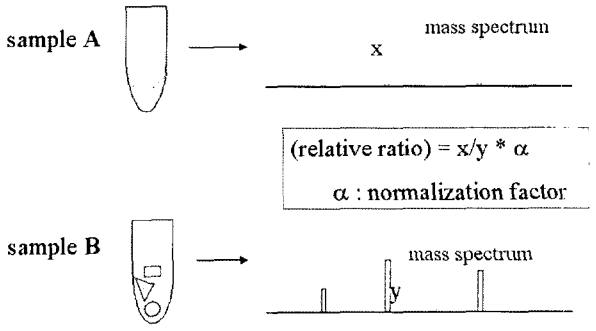


그림 2. 태그를 사용하지 않는 정량분석 방법. 두 실험에서 얻은 스펙트럼들을 정렬, 정규화 과정에 의해 분석한다. 정규화는 라벨링 방법에서의 정규화와 같이 대다수의 단백질이 양적인 차이가 없다는 가정 하에 Scatter plot 의 분석에 의해 계산된다.

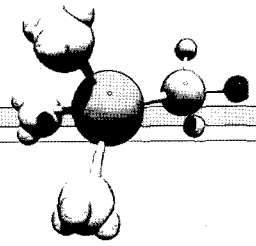
이고 보다 정확한 결과를 얻으려는 시도이다. 아직은 많은 실험 결과가 발표되지는 않았지만, 분석 장치와 데이터 해석 알고리듬의 발전과 함께 좋은 결과를 얻을 수 있으리라 예상된다.

Ⅳ. 맺음말

단백질의 정량분석 결과는 분자생물학 수준에서 생명현상을 설명하는 데에 매우 중요한 단서를 준다. 또한 환자와 정상인 사이의 단백질 분포의 차이, 환자의 질병이 진행하는 동안 나타나는 단백질 프로파일 변화의 정확한 분석은 바이오마커의 발굴을 위한 기초 자료를 제공할 수 있다.

프로테오믹스에서의 정량분석은 해마다 새로운 기법의 개발을 거듭하면서 급속한 발전을 이루고 있다. 그러나, 현재 프로테오믹스로 동정하는 단백질의 개수가 한 시료에서 최대 수천 개에 불과하며, 미량의 단백질을 높은 신뢰도로 동정하는 것은 계속적인 기술개발이 필요하다. 초기 암진단을 위한 바이오마커의 발굴과 같은 문제에 직면하면, 우리는 혈액 내에서 특정 단백질의 미세한 양적 변화와 다양한 상태변화를 감지할 수 있어야 한다. 특히 혈액 내의 단백질은 농도의 범위가 광범위하여 주요단백질(major protein)들과 미량의 단백질을 분리하여 분석하는 기술이 필요하다. 한편 단백질 프로파일에 대한 개개인의 차이를 감안하면서 질병 관련 단백질의 변화를 파악하기 위한 시료 선택 및 데이터 해석 방법도 문제로 대두된다. 단백질의 글리코실화(glycosylation), 인산화(phosphorylation) 등의 여러 가지 수식화(post-translational fication)의 양상에 따른 기능의 변화가 생체 내에 미치는 영향은 매우 클 것으로 예상되지만, 이에 대한 프로테오믹스에서의 접근은 해결해야 할 문제들을 많이 남기고 있다. 질병의 정확한 진단을 위해서는 미량의 단백질 변화를 추적하는 기술과 함께, 단백질

8) Old, W.M., Meyer-Arendt, K., Aveline-Wolf, L., Pierce, K.G., Mendoza, A., Sevinsky, J.R., Resing, K.A., Ahn, N.G., 'Comparison of label-free methods for quantifying human proteins by shotgun proteomics', *Mol. Cell. Proteomics*, 2005, 4: 1487-502.



들의 집합적인 변화 패턴의 감지, 복잡한 요인들을 포함하는 데이터의 통계학적인 분석기술이 개발되어야 한다.

이처럼 프로테오믹스에서의 정량분석을 통한 생명현상의 규명과 바이오마커의 발굴에는 첨단 초고속 분석 장비를 활용한 대량의 실험과 더불어 여러 과학분야, 기술 분야에서의 협력과 다양한 정보의 통합이 이루어져야 한다. 이는 일개 실험실, 일개 연구기관에서 감당하기에는 벅찬 분야가 아닐 수 없다. 현재는 프로테오믹스를 활용한 바이오마커 발굴을 위한 몇몇 국제 협력 프로그램들이 추진되어 각 연구 영역에서 세계적인 전문가들이 공동연구 프로젝트에 참여하고 있으며, 국내에서도 전문 연구팀들의 협력체제가 구축되고 있다. 각 분야 연구진들이 보다 활발하게 참여하고 적극적으로 협력하며 지속적인 투자를 유치하여 국제수준의 연구능력을 확보함으로써, 우리나라에서도 세계 수준의 연구 성과를 쏟아내는 미래가 오기를 기대한다.