DOI: 10.3745/KIPSTD.2008.15-D.5.681

데이터 큐브를 이용한 폐암 2-DE 젤 이미지에서의 예외 탐사

심 정 은[†] · 이 원 석[††]

요 약

단백질체학에서 특정 조건 하에서 단백질의 기능 이상 및 구조 변형 유무를 규명하고 질병 과정을 추적하는 것은 중요한 연구이다. 일반적으로 단백질의 발현량 변화 분석에는 통계적 방법이 많이 사용되고 있으며 단백질 상용 이미지 분석 소프트웨어에서 제공하는 그래픽을 이용한 방법들도 있으나, 이 방법들은 많은 조직 내에 존재하는 수많은 단백질을 수동으로 비교해야 하는 어려움이 있다. 본 논문에서는 데이터베이스와 데이터마이닝 기법을 이용하여 OLAP 데이터 큐브와 Discovery-driven 탐색의 응용 방법을 제안한다. 데이터 큐브의 특성을 이용함에 의해서, 질병에 의해 발현량이 변하는 단백질 뿐 아니라 임상적 특성과 단백질의 영향 관계를 분석하는 것이 가능하다. 데이터 큐브에서 단백질의 발현량 변화 분석에 적합한 데이터 큐브의 척도와Discovery-driven 탐색을 위한 예외 지표를 제안하고, 특히 In-exception을 계산하는데 있어서의 계산량 감소 방안을 제시한다. 실험을 통해 폐암 2-DE 데이터에서 데이터 큐브와 Discovery-driven 방법이 유용함을 보인다.

키워드 : 프로테옴 인포매틱스, 데이터마이닝, OLAP, 이차원전기영동

Discovery-Driven Exploration Method in Lung Cancer 2-DE Gel Images Using the Data Cube

Jung Eun Shim[†] · Won Suk Lee[††]

ABSTRACT

In proteomics research, the identification of differentially expressed proteins observed under specific conditions is one of key issues. There are several ways to detect the change of a specific protein's expression level such as statistical analysis and graphical visualization. However, it is quiet difficult to handle the spot information of an individual protein manually by these methods, because there are a considerable number of proteins in a tissue sample. In this paper, using database and data mining techniques, the application plan of OLAP data cube and Discovery-driven exploration is proposed. By using data cubes, it is possible to analyze the relationship between proteins and relevant clinical information as well as analyzing the differentially expressed proteins by disease. We propose the measure and exception indicators which are suitable to analyzing protein expression level changes are proposed. In addition, we proposed the reducing method of calculating *InExp* in Discovery-driven exploration. We also evaluate the utility and effectiveness of the data cube and Discovery-driven exploration in the lung cancer 2-DE gel image.

Keywords : Proteome Informatics, Data Mining, On-Line Analytical Processing, Two-Dimensional Electrophoresis

1. 서 론

단백질체학 연구는 주어진 셀이나 조직, 생물체에 표현된 단백질의 프로파일에 대한 조직적인 분석을 다룬다. 단백질체학 연구의 목적은 임의의 조직에서 어떤 단백질이 발견되는지, 특정 조건 하에서 단백질이 어떻게 상호 작용 하는 지를 분석하는 것이다. 이런 목적에서, 임의의 조직에서 특정 조건에 따른 단백질의 발현량의 변화 분석은 조직의 기능 장애를 일으키는 단백질의 도출에 있어서 핵심 이슈이다. 단백질 분석에는 Two Dimensional Electrophoresis(2-DE)와 Non-2-DE의 두 가지 기술이 사용되며, 전자는 전기영동 방식을 사용하여 조직에서 단백질을 분리한다[1]. 후자는 동위원소-코드 친화성 태그 (isotope coded affinity tag: ICAT)[2], 질량-코드 다량 태깅 mass-coded abundance tagging(MCAT)[3]과 같은 특정 친화성 태그나 액체 색층분석 질량 분광기(liquid chromatography-mass spectrometry: LC-MS)를 사용하며, 자동화에 유용하고 처리율이 높다. Non-2-DE 기술이 보다 정확한 결과를 제공하지만, 여전히 2-DE 기술이 가격 효율성 등으로 인해 단백질의 발현 패턴 분석에 주로 이용되는 기술이다[4, 5]. 그러나 2-DE의 결과는 평균적으로 1000개 이상의 스팟을 포함하는 2-DE 젤 이미지로 표현된

※ 이 논문은 2008년도 정부(과학기술부)의 재원으로 한국문화재단의 국가지정연구실사업으로 수행된 연구임(No.R0A-2006-000-10225-0).
† 준 회 원 : 연세대학교 컴퓨터과학과 박사과정
†† 종신회원 : 연세대학교 컴퓨터과학과 교수
논문접수 : 2007년 8월 8일
수 정 일 : 1차 2008년 7월 8일
심사완료 : 2008년 7월 16일

다. 각종 질병에 대한 잠재적 마커 단백질의 전체 집합을 찾기 위해서는 많은 작업이 요구된다[6]. 단백질의 발현량 변화를 검출하기 위한 가장 직관적인 방법은 눈으로 직접 2-DE 젤 이미지를 확인하는 방법이다. 또는 통계적 분석 방법[7, 8, 9]이나, 스위스 GeneBio사의 Melanie Ⅲ, 영국 Nonlinear Dynamics사의 Progenesis, 미국 Bio-rad 사의 PDQuest와 같은 상용 이미지 분석 소프트웨어 패키지에서 제공하는 그래프를 이용한 시각화 방법[10, 11]을 이용할 수 있다. 그러나, 신뢰성있는 결과를 얻기 위해서는 많은 수의 정상과 비정상 조직에서 얻은 2-DE 젤 이미지를 함께 분석할 필요가 있으며, 각 2-DE 젤 이미지에 있는 수 천개 이상의 스팟의 수를 고려할 때, 많은 수의 젤 이미지에 존재하는 각 스팟의 정보를 일일이 분석하는 것은 거의 불가능하다[12].

데이터베이스 기술은 이러한 문제를 해결하는 데 유용하다[12]. 데이터베이스 기술은 다량의 데이터를 위한 저장 구조 뿐 아니라, 이 저장 구조 위에서 데이터를 다루는 구조적인 방법도 제공하며, 데이터 웨어하우스와 OLAP(on-line analytical processing) 시스템은 다양한 사용자의 각종 요구에 대응하기 위해서 데이터에 대한 정보를 여러 형태로 구성하여 제공한다[8]. OLAP은 복잡한 분석 질의에 대한 빠른 응답을 제공하기 위한 접근 방법으로, 판매나 마케팅, 자원 관리에 관한 비즈니스 보고와 데이터 마이닝에 널리 사용되며, 더 유용한 데이터 분석을 위해서, 데이터 웨어하우스는 특정 응용 도메인의 모든 필요한 과거 데이터를 상세 분석을 목적으로 모아두는 주 정보 저장소가 된다. 데이터 웨어하우스의 이러한 특성은 수많은 단백질 중에서 특정 조건에 의해 발현량이 변하는 단백질을 도출하기 위한 논문의 목적에 부합한다. 따라서, 단백질 데이터를 OLAP 접근에 의해 효율적으로 다룬다면 유용한 지식을 얻을 수 있을 것이다. OLAP 시스템에 사용되는 다차원 데이터 모델에 있어서[13], 데이터 큐브는 차원(dimension)과 척도(measure)라는 두 요소에 의해 데이터 항목의 다양한 특성을 나타내는 데 사용된다. 차원은 데이터 큐브의 축을 이루며 계층 구조를 갖는다. 척도는 수치값으로 데이터 큐브의 각 차원을 구성하는 항목들의 조합에 해당하는 데이터들의 대표값(예를 들어, 평균)을 나타낸다. 따라서, 단백질의 발현량과 관련된 척도와 임상 정보에 대한 차원으로 구성된 데이터 큐브를 이용함에 의해 특정 질병에 대해 임의의 임상 정보와 단백질의 발현량과의 관계를 분석하는 것이 가능하다. 분석자는 몇몇 OLAP 연산에 의해 데이터 큐브를 탐색하며 예외 데이터 영역을 찾는다. 이러한 "hypothesis-driven" 탐색 방법은 차원의 수가 많아지면 데이터 큐브의 탐색 공간이 커지므로 예외 데이터를 찾는 데 한계가 있다. 그 대안으로 데이터 큐브의 탐색에 있어서 "Discovery-driven"방법이 제안되었다. 이 방법은 모든 차원의 계층에 데이터의 예외를 나타내는 이미 계산된 척도가 데이터 큐브 탐색을 안내하는 역할을 하는 것으로, 사용자가 임의의 집합 레벨에서 데이터의 비정상 패턴(예외)을 찾을 가능성을 증가시킨다[14].

본 논문에서는 폐암에 대한 2-DE 데이터베이스로부터 암세포에 의해 발현량이 변하는 단백질의 도출을 위한 데이터 큐브의 응용을 제안한다. 먼저, 스팟 데이터 집합으로부터 단백질의 발현량 변화를 나타내는 척도를 제안하며 많은 2-DE 젤 이미지로부터 동일 단백질에 대응되는 스팟들의 발현량 변화에 대한 척도와 임상 정보로 구성된 차원을 이용하여 프로테옴 데이터 큐브를 생성한다. 프로테옴 데이터 큐브를 이용함에 의해서 질병 의존 단백질 뿐 아니라 관련 임상 정보와의 관계 분석도 가능하다. 또한, 잠재적 마커 단백질을 검출하고 단백질과 임상 정보와의 관계를 분석하기 위해 데이터 큐브의 Discovery-driven 탐색 방법을 적용하는 데 있어서, 발현량 변화 양상과 정도를 나타내는 지표(indicator)를 제시하며, Discovery-driven 탐색에 있어서 보다 효율적인 탐색을 위한 방안을 제안한다. 마지막으로, 통계적 방법인 Wilcoxon nonparametric t-test와의 비교 실험을 통하여 제안한 척도와 지표의 적합성을 제시하고, 새로운 방안을 이용한 프로테옴 데이터 큐브의 Discovery-driven 탐색의 분석 효과를 실험을 통해 평가한다.

본 논문은 다음과 같이 구성된다. 2장에서는 단백질의 발현량 변화를 찾기 위해 사용되는 통계적 방법과 상용 이미지 분석 소프트웨어를 이용한 방법을 소개하고, 3장에서는 프로테옴 데이터 큐브의 모델링 방법을 보이며, 분석에 적합한 차원과 척도를 제안한다. 4장에서는 발현량 변화의 양상과 정도를 분석하기 위한 지표를 설명하고 데이터 큐브의 보다 효율적인 탐색을 위한 방안을 제시한다. 5장에서는 실험을 통해서 제안한 척도와 지표가 기존의 방법보다 폐암 2-DE 젤 이미지에서의 단백질 발현량 분석에 적합함을 보이고 마지막으로, 6장에서는 이 논문의 결론과 의의에 대해 기술한다.

2. 관련 연구

서론에서 기술한 것처럼, 단백질의 발현량 변화를 검출하기 위한 방법에는 직접 눈으로 2-DE 젤 이미지 내의 스팟을 확인하는 방법 외에 통계적 방법과 그래픽을 이용한 분석 방법이 있다. 전자는 단백질의 발현량 변화 분석에 대표적으로 사용되는 방법이고, 후자는 상용 이미지 분석 소프트웨어에서 제공하는 분석 툴을 이용한 방법이다.

통계적 방법의 주된 특징은 두 모집단 간에 차이가 존재할 확률을 구하는 것이다. 여러 샘플에서 추출한 정상 조직과 비정상 조직에 존재하는 다수의 단백질 중에서 발현량의 변화가 있는 단백질을 찾는 통계적 접근 방법에는 다음의 두 가지가 있다. 하나는 정상 조직들 내의 임의의 단백질 집합과 비정상 조직들 내의 해당 단백질 집합의 발현량 변화를 비교하는 방법이고, 다른 하나는 각 샘플에서 추출한 정상 조직 내의 임의의 단백질과 비정상 조직 내의 해당 단백질을 샘플별로 쌍으로 하여, 전체 샘플에서 해당 단백질의 발현량 변화를 비교하는 방법이다. 전자는 독립된 두 군 간의 평균 비교로, Mann-Whitney test, Sign test, Student's t-test등의 통계적 방법을 적용할 수 있다. 그러나, 단백질 데이터의 특성 상 정상 조직에 속하는 임의의 단백질에 해당하는 스팟의 집합과 비정상 조직에 속하는 해당 단백질의 스팟 집합을

평균적으로 비교하기에는 스팟 농도에 있어서 서로 다른 샘플에서 추출한 이미지 간의 변이가 매우 크므로, 이 방법은 적합하지 않다. 즉, 정상 조직에서 추출한 이미지들 내의 스팟들 간의 분산이나, 비정상 조직에서 추출한 이미지들 내의 스팟들 간의 분산이 크므로 이들을 평균적으로 비교하는 것은 부적합하다. 후자는 한 샘플에서 추출된 정상과 비정상 조직 내의 단백질을 서로 매치하고, 쌍을 이룬 각 샘플을 모두 매치함에 의해 생성된 쌍을 이룬 두 집단을 비교하는 것으로, Paired t-test방법이 있다. 매치 과정을 통해서 임의의 단백질을 나타내는 각 젤 이미지 내의 스팟들은 일관된 매치 넘버를 갖게되며, 같은 매치 넘버를 가진 스팟들의 농도값을 이용해 통계적 분석이 이루어진다. 여기서는 실험에 쓰이는 데이터가 정규 분포를 이루지 않으므로, Wilcoxon nonparametric paired t-test방법이 적합하다. 이 방법은 쌍을 이룬 단백질의 발현량 차이의 크기를 증감 별로 순위를 취하여 두 집단 간의 변화를 비교하는 것이다.

Melanie Ⅲ와 Progenesis는 잘 알려진 상용 이미지 분석 소프트웨어 패키지이다. 이것들은 발현량이 변하는 단백질의 분석을 위해 Student t-test, Wilcoxon or Mann-Whitney test, Kolmogorov test와 같은 몇몇 통계적 방법을 제공함과 함께 그 외에 다양한 그래픽적 분석 방법을 제공한다. 이론적으로 2-DE 젤 이미지의 분석에 있어서, 그래픽을 이용한 분석 방법이나 통계적 방법으로 잠재적 마커 단백질을 도출하는 것이 가능하지만, 하나의 2-DE 젤 이미지에는 평균적으로 천개 이상의 스팟이 존재하므로 개개의 스팟에 대해 일일이 분석하여 발현량이 변하는 단백질을 도출하기는 어렵다. 따라서, 많은 수의 스팟 정보를 효율적으로 다루기 위해서는 2-DE 기반의 데이터베이스 구축이 필수적이며, 데이터베이스를 이용한 새로운 분석 방법이 필요하다.

OLAP의 대표적인 데이터 모델이 데이터 큐브로 알려진 다차원 데이터베이스이다. 데이터 큐브는 데이터를 다차원으로 모델링하고 보여 지도록 하며, 테이블의 스타 스키마로부터 만들어진다[15]. 스타 스키마의 가운데에는 사실 테이블(fact table)이 있고, 사실 테이블로부터 몇몇 차원 테이블(dimension table)이 연결되어 있다. 이 테이블들은 관계 데이터(relational data)의 집합들(aggregations)을 어떻게 분석할 지를 나타낸다. 데이터 큐브에서, 데이터는 OLAP 시스템의 다차원 데이터 큐브에 요약되어 저장된다. 데이터 큐브의 계층에 따른 정보는 dril-down, roll-up, slice, dice등과 같은 OLAP 연산에 의해 탐색된다. 이와 같은 연산이 분석자로 하여금 데이터 큐브를 탐색 가능하게 하지만, 데이터 큐브에서 의미가 있는 어떤 부분에 도달하도록 지원하지는 않는다. 따라서 분석자는 자신의 가정을 기초로 데이터의 예외나 이례적인 부분을 찾는다. 이러한 데이터 큐브의 hypothesis-driven 탐색은 사용자가 데이터 큐브 내의 각 계층의 데이터를 일일이 탐색하며 예외를 찾는 것은 지루한 작업이다. 게다가 하위 레벨에 예외 데이터가 있다 하더라도 해당 레벨에서는 예외로 나타나지 않을 수도 있다. 결국 예외가 발생하더라도 분석자가 그 예외를 발견하지 못할 수도 있다[14].

"Discovery-driven" 방법[14]은 데이터 큐브의 각 레벨에서 발생하는 예외를 정의할 수 있는 미리 계산된 지표에 의해 데이터 큐브의 탐색을 지원하여, 데이터 큐브의 임의 레벨에서 데이터 예외를 도출할 수 있도록 한다. 미리 계산된 지표는 데이터 큐브의 각 셀의 값이 통계적으로 기대되는 값과 크게 차이가 나는 경우 예외를 나타낸다. 특히 차원의 수가 많거나 데이터 큐브의 계층이 복잡한 경우, 수동으로 데이터 큐브를 탐색할 때의 한계를 크게 줄일 수 있다[14].

3. 단백질의 발현량 변화 분석을 위한 데이터 큐브

3.1 2-DE 젤 이미지의 스팟 정보

2-DE는 조직의 샘플로부터 단백질을 분리하는 데 널리 사용되는 기술로, 전기 영동 방법으로 단백질을 분리하며, 그 결과는 각 단백질이 스팟으로 표현된 2차원 젤 이미지이다. 2-DE 젤 이미지 분석의 두 가지 기본적인 문제는 스팟의 검출과 매칭이다. 여기서는 이러한 문제를 Melanie나 Progenesis와 같은 2-DE 이미지 분석 소프트웨어를 사용하여 해결한다[10, 11].

2-DE 젤 이미지 분석 소프트웨어를 통해 2-DE 젤 이미지로부터 스팟을 검출하면, 하나의 2-DE 젤 이미지 당 몇천개의 스팟이 검출된다. 각 스팟은 하나의 단백질에 대응되며, 유일한 spotID가 부여된다. 또한 젤 이미지 내에서 각 스팟을 나타내는 정보로, 좌표값 (x, y)와 농도 정보 Od(Optical Density), Vol(Volumn), %Od, %Vol 값이 생성된다. %Od와 %Vol은 Od와 Vol을 표준화한 값으로, 젤 이미지 내의 모든 스팟의 Od, Vol의 합을 100으로 했을 때 각 스팟의 Od, Vol 값의 백분율을 나타낸 값이다. 스팟의 농도 정보는 개개의 이미지가 어떻게 생성되느냐에 따라 값이 달라질 수 있으므로, 하나의 젤 이미지 내의 스팟과 다른 젤 이미지 내의 스팟을 비교하기 위한 표준 척도로는 적합하지 않다. 따라서 표준화한 %Od나 %Vol 값을 사용하며, 이 값은 해당 스팟에 대응되는 단백질의 발현량을 나타내는 값으로 쓰인다.

[정의 1] 단백질의 발현량

2-DE 젤 이미지 I 내의 m개의 스팟이 주어졌을 때,

$$S_I = \{s_k | s_k \text{는 이미지 } I \text{내에 존재하는 스팟}, 0 \le k \le m\}.$$

스팟 s_k가 나타내는 단백질의 발현량은 다음과 같이 나타낸다.

$ExpLevel(s_k)$ = %Vol 또는 %Od (스팟 s_k의 속성).

□

젤 이미지와 각 스팟 리스트가 주어졌을 때, 각 젤 이미지 내에서 동일한 단백질에 대응하는 스팟들을 찾기 위해 매칭 과정이 수행된다. 젤 이미지의 집합에 존재하는 스팟들은 하나의 참조 젤 이미지(reference gel image) 내의 주어진 스팟과 매치되어 pairing class를 이룬다. pairing class는 젤 이미지 간에 단백질의 발현량 변화를 찾고 분석하는 데 기본 요소가 되므로, 스팟의 매칭은 2-DE 젤 이미지 분석의 주요 작업이다. 만약 참조 젤 이미지 내의 임의의 스

팟과 어떤 이미지 내의 스팟이 매치되었다면, 그 두 스팟은 pair relation에 있다고 하며, Pairing Class는 다음과 같이 정의한다.

[정의 2] Pairing Class

N개의 젤 이미지 집합 $I=\{I_1, I_2, ..., I_n\}$ 이 주어졌을 때, $I_R \in I$을 참조 젤 이미지라 하면, S_{I_R}내의 스팟은 $S_{I_i(1\le i\le N)}$ 내의 스팟들과 쌍을 이룬다. 만약, $s_r \in S_{I_R}$과 $s_j \in S_{I_i}$가 pair relation에 있다면, 이것을 $(s_r, s_j) \in P$로 나타낸다. 그리고, 참조 젤 이미지에서 spotID가 r인 스팟의 Pairing Class는 다음과 같이 나타낸다.

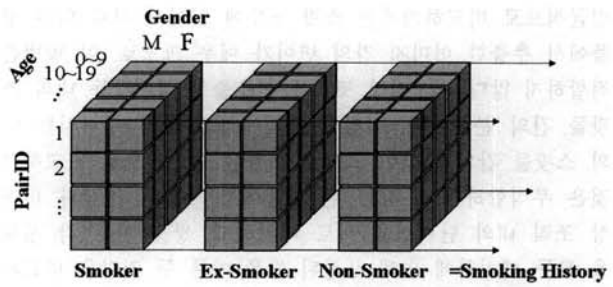
$$PC_r = \{s_j \mid (s_r, s_j) \in P, s_j \in S_{I_i}, s_r \in S_{I_R}\}$$

여기서, r은 PairID로 정의한다. □

(그림 1)은 실제 2-DE 젤 이미지에서의 pairing class의 예를 보여준다. 각 이미지 내에서 스팟의 발현 패턴을 비교해 볼 때, 스팟 A는 상대적으로 유사한 위치에 발현되었다. 이 때, 각 이미지에서 스팟 A들는 동일한 단백질을 의미하게 되며, 각 이미지에서의 이 스팟 A들의 집합을 pairing class라 할 수 있다.

3.2 데이터 큐브의 모델링

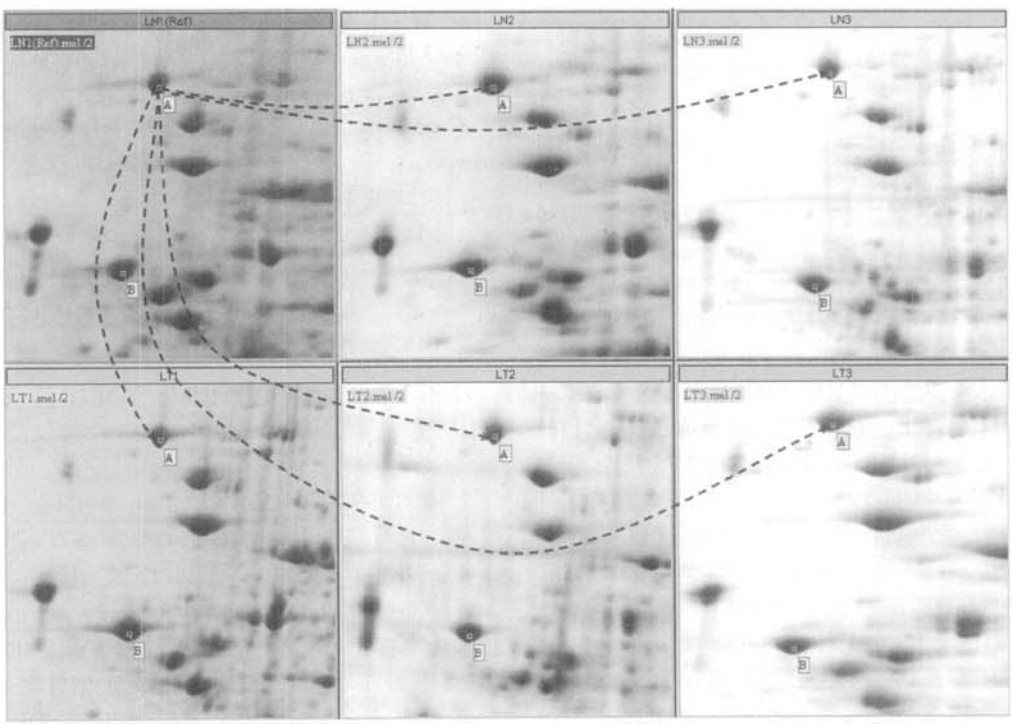
단백질체와 관련된 데이터는 질병과 관련된 환자에 대한 임상 정보와 그 환자의 질병 관련 조직으로부터 얻은 2-DE 젤 이미지 내의 스팟들의 정보로 구성된다. 이들 정보 간의



(그림 2) 데이터 큐브의 예

관계를 분석하여 비정상적이거나 예외적인 패턴을 찾기 위해 해당 데이터들이 수집되며, 여기서는 데이터가 다차원으로 모델링되고 보여지도록 데이터 큐브를 사용한다. 데이터 큐브는 차원과 척도로 정의된다. 일반적으로, 차원은 유지하고자 하는 레코드와 관련된 요소나 관점을 말하며, 데이터 큐브의 축을 이룬다[16]. 예를 들어, 단백질체 데이터 웨어하우스 내의 데이터와 관련된 레코드를 유지하기 위해서, 차원은 성별, 나이, 흡연, 음주, 병력 등의 임상 정보와 개개 단백질을 나타내는 PairID로 구성될 수 있다. 이들 차원은 단백질들이 포함된 조직의 샘플에 대한 임상 정보와 각 단백질을 구별할 정보를 유지한다. 각 차원은 그 차원과 관련된 테이블을 가지며, 이것을 차원 테이블이라 한다[16]. (그림 2)은 PairID와 성별, 나이, 흡연에 대한 임상 정보로 구성된 4차원 데이터 큐브를 나타낸 것이다.

다차원 데이터 모델은 예를 들어 '단백질의 발현량 변화'와 같은 주제를 중심으로 구성된다. 이 주제는 사실 테이블에 의해 표현되며, 사실 테이블은 사실의 이름이나 척도 및



(그림 1) Pairing Class의 예

관련 차원 테이블들의 각 키값을 포함한다. 사실 테이블에 포함되는 척도는 데이터 큐브의 셀 내의 값이 되며, 차원들 간의 관계를 분석하기 위한 수치 값이다[16]. 질병에 의한 발현량의 변화를 분석하는 데이터 큐브의 척도는 질병의 영향이 있는 조직과 없는 조직 내의 단백질의 발현량 차이를 나타내는 값이 된다. 질병에 의한 단백질의 발현량 변화를 비교하기 위해서, Pairing Class 내의 스팟들을 정상 젤 이미지 내의 스팟들로 구성된 하나의 집합으로 하고, 비정상 젤 이미지 내의 스팟들로 구성된 다른 하나의 집합으로 나눈다. 그러나 정상과 비정상 그룹을 평균적으로 비교하기에는 젤 이미지들 간에 변이가 크므로 부적합하다. 따라서 동일 환자로부터 얻은 정상과 비정상 젤 이미지 내의 스팟을 쌍으로 하여 비교하기 위해 Pairwise Pairing Class를 정의한다.

[정의 3] Pairwise Pairing Class

정상 조직의 젤 이미지 집합 $I_N = \{I_{1N}, I_{2N}, ..., I_{nN}\}$ 과 비정상 조직의 젤 이미지 집합 $I_T = \{I_{1T}, I_{2T}, ..., I_{nT}\}$ 이 주어졌을 때, 참조 젤 이미지를 $I_R \in (I_N \cup I_T)$ 이라 하자. $I_{iN} \in I_N$을 i번째 환자의 정상 조직의 젤 이미지라 하고, $I_{iT} \in I_T$를 i번째 환자의 비정상 조직의 젤 이미지라 할 때, $S_{I_{iN}}$을 I_{iN} 이미지에 존재하는 스팟들의 집합이라 하고 $S_{I_{iT}}$를 I_{iT}이미지에 있는 스팟들의 집합이라 정의하면, 참조 젤 이미지에서 spotID가 r인 스팟의 Pairwise Pairing Class는 다음과 같이 나타낸다.

$$PPC_r = \{(s_j, s_k) | s_j \in PC_r, s_k \in PC_r, s_j \in S_{I_{iN}}, s_k \in S_{I_{iT}}, s_r \in S_{I_R}\}$$

□

3.3 데이터 큐브의 차원과 척도

차원은 각 요약 레벨을 지정하는 계층과 관련되어 있다. 예를 들어, 임상 정보를 나타내는 차원의 계층 구조는 다음의 (그림 3)와 같이 나타낼 수 있다.

제안하는 데이터 큐브는 임상 정보를 포함하기 때문에 질병에 의한 발현량 변화 뿐 아니라 임상 특성에 따른 영향도 분석 가능하다. 예를 들어, 임의의 단백질이 폐암에 의해 발현량이 증가하기도 하고 감소하기도 한다면, 폐암에 대한 마커 단백질로 판단하기 어려우나, Smoker에서는 증가하고 Non-Smoker에서는 감소했다면 임상(흡연)과 관련한 의미있는 분석이 가능하다.

기존의 통계적 방법은 정상 단백질 발현량과 비정상 단백질 발현량의 차(delta)를 이용하므로, 외부 변화 요소에서 오는 단백질의 발현량과 관계 없는 미약한 발현량 차이도 단백질 발현량의 변화로 취급하게 된다. 그러나, 이러한 미약한 차이를 배제하기 위해서 본 논문에서 데이터 큐브의 척도는 발현량 변화를 차이가 아닌 배수를 이용하며, 다음과 같이 정의한다.

[정의 4] 단백질의 발현량 변화

$(s_j, s_k) \in PPC_r$에 대해서, s_j를 정상 조직의 젤 이미지 내의 스팟이라 하고, s_k를 비정상 조직의 젤 이미지 내에 있는 스팟이라 할 때, PairID r에 해당하는 임의의 단백질의 발현량 변화를 다음과 같이 정의한다. (앞으로 기호 s는 스팟 s가 나타내는 단백질을 의미하는 것으로 한다.)

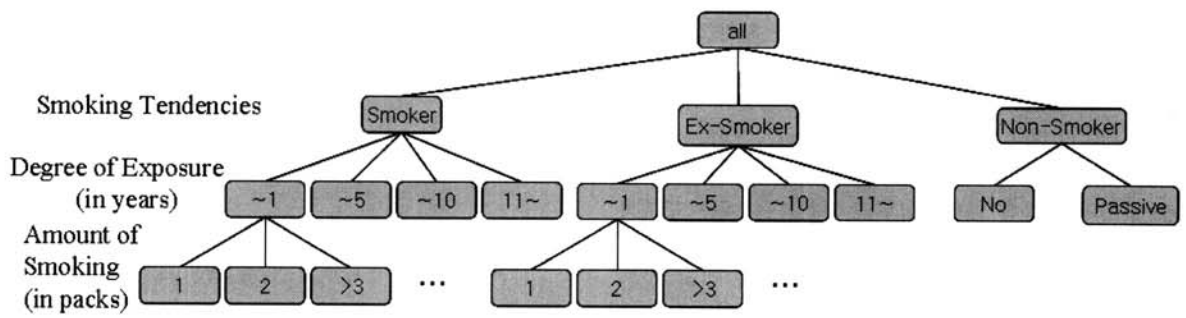
$$R\_ExpLevel(s_j) = \frac{ExpLevel(s_k)}{ExpLevel(s_j) + ExpLevel(s_k)}$$

□

프로테옴 데이터 큐브의 분석 목표는 질병에 의해 발현량이 증거하거나 감소하는 단백질을 도출하는 것이다. *R_ExpLevel(s)*는 0에서 1사이의 값을 가지며, 정상 조직 내의 단백질의 발현량과 비교하여 *R_ExpLevel(s)>0.5*이면 비정상 조직 내의 단백질의 발현량이 증가했음을 나타내고, *R_ExpLevel(s)<0.5*이면 비정상 조직 내의 단백질의 발현량이 감소했음을 나타낸다. 그러나, 단백질의 발현량과는 관계가 없는 외부 변화 요소에 의해 젤 이미지 간의 변이가 존재하므로, 단백질 발현량의 미미한 변화는 질병에 의한 변화로 보기 어렵다. 따라서 정상 조직내의 단백질의 발현량과 비교하여, 암 조직 내의 단백질의 발현량이 x배 이상이면 발현량이 증가했다고 판단하고, $1/x$배 이하이면 발현량이 감소했다고 판단한다.

[정의 5] 단백질 발현량의 변화 양상

단백질 발현량의 변화는 *R_ExpLevel(s)*의 값이 $\frac{x}{x+1}$ 보다 크면 발현량이 증가했다고 판단하고, $\frac{1}{x+1}$ 보다 작으면 감소했다고 판단한다.



(그림 3) 흡연 차원의 계층 예

- s : Under-expressed protein. if, $R\_ExpLevel(s) \le \frac{1}{x+1}$.
- s : General protein. if, $\frac{1}{x+1} < R\_ExpLevel(s) < \frac{x}{x+1}$.
- s : Over-expressed protein. if, $R\_ExpLevel(s) \ge \frac{x}{x+1}$.

□

실제로 생물학적 분석 실험에 있어서 25%이내의 발현량 변화는 생물학적으로 중요하지 않다고 판단하며 발현량의 차이가 1.5배, 2배, 또는 3배 이상인 경우를 발현량에 변화가 있다고 판단한다. 데이터 큐브의 각 셀은 다수의 pairwise pairing class들의 집계값인 척도로 표현되며 다음과 같이 정의할 수 있다.

[정의 6] 단백질의 발현량 변화를 분석하기 위한 데이터 큐브의 척도

n차원의 데이터 큐브에 대해서, l번째 차원 $d_l(1 \le l \le n)$의 i_l 위치의 척도를 다음과 같다.

- $U\_cnt_{i_1 i_2 ... i_n}$: $i_1 i_2 ... i_n$위치의 셀에서 Under-expressed protein의 수.
- $G\_cnt_{i_1 i_2 ... i_n}$: $i_1 i_2 ... i_n$위치의 셀에서 General protein의 수.
- $O\_cnt_{i_1 i_2 ... i_n}$: $i_1 i_2 ... i_n$위치의 셀에서 Over-expressed protein의 수. □

4. Discovery-driven 탐색을 이용한 질병 의존 단백질의 도출

데이터 큐브를 생성하기 위해서는 선택된 차원과 척도에 대한 데이터가 필요하다. 3장에서 설명한 요소로 데이터 큐브를 생성하고, 데이터의 임의 집합 레벨에서 예외 데이터를 찾을 기회를 증대시키기 위해 예외 지표를 계산하여 Discovery-driven 탐색이 가능하도록 한다. 이 장에서는 데이터 큐브의 예외 지표를 정의하고 이것을 이용한 Discovery-driven 탐색 과정을 기술한다. 본 논문에서 데이터 큐브는 질병 의존 단백질의 도출 뿐 아니라 해당 단백질의 임의의 임상 정보에 의한 영향을 분석하기 위한 목적으로 생성하는 것이므로, 데이터 큐브의 차원은 질병과 관계가 있을 것으로 보이는 선택된 임상 정보가 될 것이다. 각 단백질에 대한 분석이므로 개개의 단백질을 나타내는 PairID는 데이터 큐브의 기본적인 차원이 된다. PairID와 임상과 관련한 선택된 차원으로 구성된 데이터 큐브는 의미 있는 단백질을 도출하기 위해 탐색된다. 탐색은 기본적으로 데이터 큐브 차원의 최상위 레벨에서 시작한다. 데이터 큐브의 탐색에는 "Roll-up", "Drill-down", "Slice and dice", "Pivot" 등의 연산이 사용된다[16].

계층 구조의 최상위 레벨에서부터 분석자는 계층 구조의 하위 레벨로 drill-down하며 예외 데이터를 찾기 위해 탐색을 수행한다. 만약 임의의 패스를 통한 탐색이 의미 있는 결과를 찾지 못하면, 분석자는 해당 패스를 roll-up하여 다른 패스로 탐색을 시작하고, roll-up의 결과 최상위 레벨에 다다르면 다른 차원으로 drill-down하여 탐색을 계속한다. 위와 같은 "hypothesis-driven" 탐색의 대안이 모든 집합 레벨에서 데이터의 예외를 나타내는 지표를 미리 계산하여 분석자의 데이터 분석 과정을 지원하는 "Discovery-driven" 탐색 방법이다. 이 방법은 분석자가 예외 패턴을 찾을 확률을 높인다[14]. 모든 집합 레벨에서 데이터 분석 과정을 지원하는 지표를 예외 지표(exception indicator)라 하며[14], 단백질의 발현량 변화를 분석하기 위한 데이터 큐브에서 적용할 수 있는 예외 지표를 다음과 같이 정의한다.

[정의 7] 단백질의 발현량 변화를 분석하는 데이터 큐브에서 예외 지표

n차원의 데이터 큐브에서, l번째 차원 $d_l(1 \le l \le n)$의 i_l 위치의 셀의 예외 지표는 다음과 같이 정의한다.

$$U\_G\_O\_ratio_{i_1 i_2 ... i_n} = \max(U\_ratio_{i_1 i_2 ... i_n}, O\_ratio_{i_1 i_2 ... i_n}).$$

여기서,

$$U\_ratio_{i_1 i_2 ... i_n} = \frac{U\_cnt_{i_1 i_2 ... i_n}}{U\_cnt_{i_1 i_2 ... i_n} + G\_cnt_{i_1 i_2 ... i_n} + O\_cnt_{i_1 i_2 ... i_n}}$$

이고,

$$O\_ratio_{i_1 i_2 ... i_n} = \frac{O\_cnt_{i_1 i_2 ... i_n}}{U\_cnt_{i_1 i_2 ... i_n} + G\_cnt_{i_1 i_2 ... i_n} + O\_cnt_{i_1 i_2 ... i_n}}$$

이다.

즉, 예외 지표는 해당 셀 내의 모든 단백질의 수에 대한 Under-expressed 또는 Over-expressed 단백질의 수의 비율로 나타낸다. □

Discovery-Driven 방법은 예외 지표를 이용하여 데이터 큐브를 탐색하는 데 있어서, 다음의 세 가지 값을 사용한다[14].

- *SelfExp*: 현재 레벨에서 어떤 셀이 예외인지를 나타내는 값이다. 임의의 셀의 예외 지표 값과 정해진 임계값에 따라 정의된다. 즉, 예외 지표의 값이 임의의 임계값 τ보다 크면 해당 셀은 예외로 판단한다. 앞 절에서 정의한 예외 지표에 대해서는 다음과 같이 정의된다.

$$SelfExp(U\_G\_O\_ratio_{i_1 i_2 ... i_n}) = \max(U\_G\_O\_ratio_{i_1 i_2 ... i_n} - \tau, 0).$$

- *InExp*: 현재 셀에서 drill-down 연산을 통해 도달할 수 있는 모든 셀들의 예외 정도를 나타내는 값이다. 앞 절에서 정의한 예외 지표에 대해서는 다음과 같이 정의된다.

$$InExp(U\_G\_O\_ratio_{i_1 i_2 ... i_n}) = \max\{SelfExp(U\_G\_O\_Ratio_{j_1 j_2 ... j_n}) \mid (\forall l, 1 \le l \le n, i_l = j_l \text{ or } i_l = +) \,\&\, (\{j_1, ..., j_n\} \ne \{i_1, ..., i_n\})\}.$$

- *PathExp*: 현재 셀에서 가능한 각 drill-down 패스들

중에서 특정 패스로 drill-down했을 때 기대되는 예외 정도를 나타내는 값이다. 앞 절에서 정의한 예외 지표에 대해서는 다음과 같이 정의된다.

$$PathExp(U\_G\_O\_ratio_{i_1 i_2 \cdots i_n}, k) = \max\{SelfExp(U\_G\_O\_ratio_{j_1 j_2 \cdots j_n}) \mid (\forall l, 1 \le l \le n, i_l = j_l \text{ or } i_l = +) \,\&\, j_l \neq +\}, \forall k \text{ where } i_k = +.$$

*SelfExp*의 값은 의미있는 셀을 도출하도록 하며, *InExp*과 *PathExp*은 데이터 큐브를 효율적으로 탐색할 수 있도록 한다[14]. *InExp*과 *PathExp*은 하위 레벨에 해당하는 모든 셀의 예외 지표를 계산해야 하므로 계산량이 많다. 그러나 현재 레벨에서의 셀의 값과 *SelfExp*의 임계값, 그리고 환자의 임상 정보 분포를 고려하여 그 계산량을 줄일 수 있다. 계산량을 줄이는 방안은 다음 두 가지로 생각해 볼 수 있다. 먼저, 현재 셀이 *SelfExp*인 경우에 해당 셀이 나타내는 발현량의 변화 경향이 하위 레벨로 drill down하더라도 변화가 없을 경우, 즉 drill down해도 현재 셀이 나타내는 의미 이상의 정보를 얻을 수 없는 경우이다. 두번째는 현재 셀이 *SelfExp*이 아닌 경우에, 해당 셀 내의 단백질 중 General protein의 비율이 높아서 하위 레벨로 drill down하더라도 U_ratio의 값이나 O_ratio의 값이 예외가 될 가능성이 없는 경우이다. 만약 현재 셀이 나타내는 단백질이 Under-expressed 이거나 Over-expressed 단백질로 판단되고 그 예외 지표의 값이 1이라면, 더 이상의 drill-down연산이 의미가 없으므로, *InExp*의 계산은 필요 없게 된다. 또한 drill- down할 차원의 환자 분포를 알고 있으면, 현재 셀을 drill-down해도 하위에 다른 의미 있는 결과가 도출되지 않는다는 상한을 정할 수 있을 것이다.

- 현재 *SelfExp*인 경우,

$$\min(U\_ratio,\ O\_ratio) \ge (the\ ratio\ of\ minimum\ attribute \times \tau).$$

- 현재 *SelfExp*이 아닌 경우,

$$\max(U\_ratio,\ O\_ratio) \ge (the\ ratio\ of\ minimum\ attribute \times \tau).$$

위 식에서 minimum attribute는 차원의 하위 항목 중 최소 샘플을 포함한 항목을 뜻한다. 예를 들어, 전체 환자 샘플의 수가 100개라고 할 때, 성별의 하위 항목인 남성에 해당하는 샘플이 70개이고 여성에 해당하는 샘플이 30개라면, minimum attribute는 여성이며, 그 비율은 0.3이다. τ는 예외 지표의 임계값이므로, drill down의 대상 차원의 하위 항목들의 분포에 따라 drill down을 수행하여 예외 상황이 발생할 가능성이 있는 셀에 대해서만 *InExp*의 계산을 수행한다.

5. 실험 및 결과 분석

실험에서는 폐암 환자 20명의 데이터를 각각 이용했다. 환자의 질병 관련 조직에서 얻은 정상 조직과 비정상 조직으로부터 2-DE 젤 이미지를 생성하고, 이미지 분석 소프트웨어(Progenesis)를 이용하여 실험 데이터를 추출하였다.

실험에 쓰인 폐암 관련 환자의 임상 정보 별 분포는 다음 <표 1>과 같다. 폐암 데이터의 데이터 큐브는 성별, 흡연, PFT (Pulmonary Funtion Test)정보를 차원으로 하였다. PFT는 환자의 임상 정보 중 폐기능 검사에서 1초간 노력성호기량(FEV1)의 노력폐활량(Forced Vital Capacity: FVC)에 대한 비율값으로 FEV1/FVC를 계산하여, 80%이상인 경우는 restriction, 70~80%의 값을 갖는 경우는 normal, 70%이하 인 경우는 obstruction을 나타낸다. 폐암의 경우 obstruction인 경우와 그렇지 않은 경우에 있어 비교 분석의 의미가 있다. 임상 정보가 입력되지 않은 경우 Null로 처리하였다. 위의 데이터 집합은 스팟 검출과 매칭 과정으로부터 1538개의 PairID가 생성되었다.

5.1 척도와 예외 지표의 적합성

본 절에서는 단백질 발현량 분석에 대표적으로 사용되는 통계적 분석 방법 중 데이터의 특성에 가장 적합한 Wilcoxon nonparametric paired t-test 방법을 논문에서 제안한 예외 지표와 비교 실험하여, 기존 방법의 문제점을 제시하고 제안한 척도와 예외 지표가 단백질 발현량 변화 분석에 적합함을 보인다. 통계적 방법의 분석은 통계 분석툴인 SPSS v11.0을 이용하였다.

단백질의 발현량을 나타내는 값으로는 %Vol 값을 이용하였다. 실제 실험에서는 단백질의 발현량을 나타내는 %Vol의 값이 정상과 비정상 조직의 젤 이미지에서 모두 0.1 이하인 스팟은 단백질로 고려하지 않고 실험 과정상에 발생한 오류(노이즈)로 보고, 필터링을 하였다. 실험은 존재하는 1538개의 단백질 중 7개 이상의 환자 조직에서 쌍으로 나타난 단백질 114개에 대해서 수행하였다. 즉, PPC_r의 크기가 7이상인 PairID를 실험 데이터로 하였다. 기존의 통계적 방법에는 다음의 두 가지 문제점이 있다. 통계적 방법에서는 정상과 암 조직 내의 단백질 발현량 비교에 쓰이는 기준이 발현량이 차이(delta)이므로, 실험 상의 외부 변화 요소에서 오는 단백질의 발현량과 관계없는 미약한 발현량 차이도 변화로 취급하는 문제점이 있다. 그러나, 미약한 차이를 배제하기 위해서 해당 샘플을 필터링하는 것은 전체 데이터의 특성을 잃게 되는 것이므로 부적절하다. 따라서 본 논문에서 데이터 큐브의 척도는 발현량 변화를 차이가 아닌 배수를 이용

〈표 1〉 폐암 데이터의 임상 정보 분포

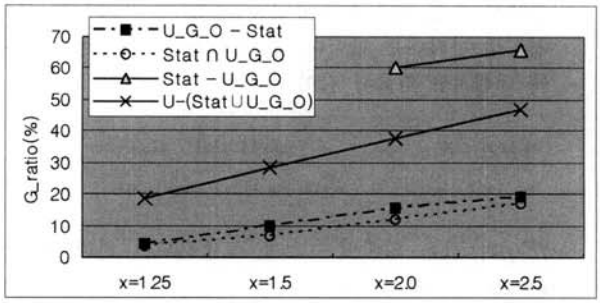
Sex	Male				Female	Null
PFT	Restriction	Normal	Obstruction	Null	Restriction	Null
Smoke	3	1	3	1	0	0
Ex-Smoke	0	0	0	1	0	1
Non-Smoke	1	0	2	0	3	0
Null	0	1	1	1	0	1

한다. 또한, 2-DE 실험의 특성 상 실험 샘플들 간의 변이가 크다는 점을 고려하지 않고 단지 발현량의 차이의 크기로 순위를 취하여 통계 값을 구하므로, 다수의 환자 샘플에서 감소하는 경향을 보이는 단백질 발현량이 소수의 샘플에서 큰 차이로 증가하는 경우, 두 경향이 상쇄됨에 의해 해당 단백질은 발현량이 변하는 단백질로 판단되지 않는 문제점이 있다. 본 실험에서는 정의 5에서 발현량의 차이를 나타내는 임계값 $x>=1.25$ 로 하여 1.25배 이상의 차이를 변화가 있다고 고려했다.

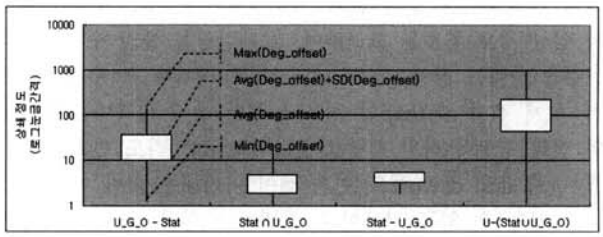
다음의 실험들에서 (Stat-U_G_O)는 통계적 방법의 95% 신뢰 수준에서만 발현량이 변한다고 판단되는 단백질을, (U_G_O-Stat)는 제안하는 방법에서만 발현량이 변한다고 판단되는 단백질을 나타낸다. (Stat∩U_G_O)는 두 척도에서 모두 발현량이 변한다고 판단되는 단백질을 (U-(Stat∪U_G_O))는 어느 척도에서도 발현량의 변화가 없다고 판단되는 단백질을 나타낸 것이다. (Stat∩U_G_O)와 (U-(Stat∪U_G_O))는 비교 대상 두 척도에 대해 판단이 일치하는 데이터로 별도의 분석이 요구되지 않으나, (Stat-U_G_O)와 (U_G_O-Stat)는 두 척도에 따른 판단이 상이하므로 비교 분석의 필요가 있다.

(그림 4)은 통계적 방법의 첫 번째 단점인 미약한 단백질 발현량 차이를 변화로 인식한 문제점을 나타낸 그래프이다. 발현량의 차이를 나타내는 정의 5의 임계값 x를 변화하면서, 전체 단백질 중에서 General Protein으로 분류되는 단백질의 비율의 평균치를 나타낸 것으로, 네 가지 계열 모두 x값이 증가함에 따라 그 비율이 증가함을 알 수 있다. 그림 4에서 $x=1.25$와 $x=1.5$에서는 (Stat-U_G_O)계열의 데이터가 없음을 나타낸다. (U-(Stat∪U_G_O))계열과 분석의 대상이 되는 (Stat-U_G_O)계열의 경우, 그 비율이 다른 계열의 데이터에 비해 큼을 알 수 있다. 즉, 실제 질병의 영향에 의한 발현량의 변화로 보기 어려운 미약한 변화가 배수를 이용한 척도에서는 General protein으로 판단되었는데 반해, 통계적 방법에서는 발현량의 변화로 취급된 것이다.

(그림 5)는 통계적 방법의 두 번째 문제점을 보이기 위한 것으로, (U_G_O-stat)에 해당하는 데이터가 통계적 방법에서는 의미 있는 데이터로 판단되지 않는 이유를 나타낸 것이다. 본 실험에서 $x=2.0$인 경우를 나타내었다. 그래프의 y축은 발현량 증감의 상쇄 정도를 로그 눈금 간격으로 나타낸 것이며 상쇄 정도를 나타내는 계산식은 아래에 제시하였다. 그래프는 각 계열 별로, 해당 계열에 해당하는 PairID들에 대해 상쇄 정도의 최소값(Min), 최소값(Max), 평균(Avg), 표준편차(SD)를 이용하여 나타내었다. 하나의 PairID r에 대



(그림 4) 통계적 방법의 미약한 발현량 변화 인식 문제



(그림 5) 통계적 방법의 상쇄 효과

해 상쇄 정도를 나타낸 척도는 다음과 같다.

$$Deg\_offset(r)=\frac{\sum_{PairID=r\&(s_j,s_k)\in PPC_r}|Degree(s_j,s_k)|}{\sum_{PairID=r\&(s_j,s_k)\in PPC_r}Degree(s_j,s_k)},$$

$$Degree(s_j,s_k)=\begin{cases}\frac{ExpValue(s_j)}{ExpValue(s_k)} & \text{if } ExpValue(s_j)>ExpValue(s_k)\\ -\frac{ExpValue(s_k)}{ExpValue(s_j)} & \text{if } ExpValue(s_j)<ExpValue(s_k)\\ 0 & \text{otherwise}\end{cases}$$

즉, PPC_r내의 단백질의 발현량 변화 경향이 일정하면 *avg_sum*과 *avg_delta*의 값에 큰 차이가 없으므로, *Deg_offset*의 값이 작지만, 단백질이 증가하기도 하고 감소하기도 하는 경우에는 *Deg_offset* 계산식의 분모에 해당하는 *avg_sum*은 커지고, 분자에 해당하는 *avg_delta*는 0에 가까운 값이 되므로, *Deg_offset*의 값이 커지는 결과가 된다.

(그림 5)는 (U-(stat∪U_G_O))와 (U_G_O-stat)의 상쇄 정도가 큼을 보여준다. 즉, 통계적 방법에 따른 경우, 소수의 경향에 의해 다수의 경향이 상쇄되어 의미 있는 단백질을 의미 있다고 판단하지 않는 경우가 발생함을 나타낸다. 따라서 증감 별로 정도에 따른 순위를 이용하는 방법은 부적합하며, 제시한 척도와 같이 증감 별로 비율을 파악하여 경향을 분석하는 것이 적합하다. <표 2>는 기존의 다양한 단백질 발현량 분석 기법들의 장단점을 비교하였다.

〈표 2〉 단백질의 발현량 분석 기법들의 비교

	그래픽 분석 기법	통계적 분석 기법	예외탐사 기법
장점	- 2DE 젤 이미지 내에 포함된 오류 및 변형을 분석자가 바로 인지하여 분석이 가능	- 기존의 통계분석 툴을 이용할 수 있음 - 이미 검증된 분석 방법임	- OLAP 시스템에 탑재하여 자동화된 분석 방법을 제공 - 발현량 변화를 차이가 아닌 배수를 이용하기 때문에 상대적인 발현량 변화 비교가 가능함 - over 또는 under의 개수 비율로 경향을 분석하기 때문에 통계적 방법에서 발생하는 상쇄효과 문제가 발생하지 않음
단점	- 분석자가 수동으로 분석해야 하기 때문에 많은 시간과 노력이 소요됨	- 상대적으로 미약한 발현량의 변화를 강하게 인지할 수 있음 - 하나의 큰 발현변화가 다수의 작은 발현변화를 상쇄시킬 수 있음	

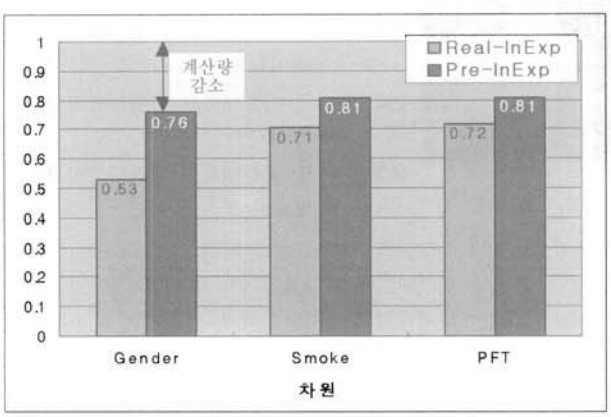
5.2 InExp의 계산량 감소

이 실험은 InExp의 계산량 감소 방안의 효율성에 대한 실험으로, 폐암 데이터 집합에 존재하는 1538개의 단백질 중 7개 이상의 환자 조직에서 쌍으로 나타난 단백질 612개에 대해서 수행하였다. 즉, %Vol의 값에 관계없이 PPC_r의 크기가 7이상인 총 612개의 PairID를 실험 데이터로 하였다. (그림 6)는 발현량 변화의 임계값 x=1.5로 하고, *SelfExp*의 임계값 τ를 0.5로 하였을 때, *InExp*계산량에 대한 감소 효과를 나타낸 그래프이다. 계열 Pre-InExp는 현재 데이터 큐브의 셀에 대해서 *InExp*을 계산할 대상 셀의 비율을, Real-InExp 계열은 실제 *InExp*인 셀의 비율을 나타낸다. 실험에서 차원 Gender, Smoke, PFT에 대한 InExp을 계산하기 위해 모든 셀을 탐색할 필요 없이 전체 셀의 75~80%의 셀만을 탐색하여 20% 이상의 계산량을 감소시킬 수 있었다.

5.3 데이터 큐브의 탐색

(그림 7)은 앞 절의 실험에서 사용한 폐암 데이터로 PairID에 의한 1차원 데이터 큐브의 일부를 나타낸 것이다. 큐브의 셀에 나타낸 척도는 U_cnt, G_cnt, O_cnt이며, 정상과 비정상 조직에서의 단백질이 1.5배 이상 발현량 변화가 있는 경우를 변화가 있다고 판단하였다. 그림에는 예외 지표의 임계값 τ를 0.5로 하였을 때의 *SelfExp*을 나타내었다. 이 그림에서, 703, 1210, 1313, 1450, 1574번 PairID는 Under-expressed protein으로 판단되었고, 879, 1064, 1306번 PairID는 Over-expressed protein으로 판단되었음을 나타낸다.

(그림 8)은 위와 같은 데이터 큐브에서 각 차원 별 *InExp*을 나타낸다. 데이터 큐브의 제시한 일부만 보았을 때는 성별, 흡연, PFT 차원으로 drill-down하는 것이 하위 레벨에 더 의미 있는 셀이 많은 것을 알 수 있다. 위 실험에서는 Under-expressed인지 Over-expressed인지의 판단만을 나타냈으나, 음영의 밝기나 색을 달리 하여 예외의 정도를 나타낼 수도 있으며, 정의 5의 x값이나 예외 지표의 임계값 τ를 변화하여 보다 상세한 분석이 가능하다. 이와 같은 데이터 큐브의 상세 분석을 통해 수 백 여 개의 단백질 데이터 중에 생화학적인 추가 분석에 의미가 있다고 생각되는 단백질



(그림 6) InExp 계산량의 감소

PAIR_ID	U_G_O_Count
703	11_4_5
879	4_3_11
926	8_0_8
954	9_5_4
1064	1_4_11
1210	11_7_2
1306	3_3_14
1313	11_3_6
1450	11_3_4
1477	8_1_8
1574	10_5_1
1585	3_6_8
1809	8_3_6
1818	7_5_4

: Under-Expressed
: Over-Expressed

(그림 7) 1-D 데이터 큐브

PAIR_ID	U_G_O_Count
703	11_4_5
879	4_3_11
926	8_0_8
954	9_5_4
1064	1_4_11
1210	11_7_2
1306	3_3_14
1313	11_3_6
1450	11_3_4
1477	8_1_8
1574	10_5_1
1585	3_6_8
1809	8_3_6
1818	7_5_4

: InExp - Gender dimension
: InExp - PFT dimension
: InExp - Smoke dimension

(그림 8) InExp의 표시

들의 우선순위를 제시할 수 있으며, 데이터 큐브의 탐색 연산에 의해 임상 정보와 질병 의존 단백질 간의 상관관계도 분석 가능하다. 본 논문에서 제안한 방법은 많은 조직 내에 존재하는 수 천 개 이상의 단백질을 임상 정보와 연계하여 보다 체계적이고 효율적으로 분석할 수 있으며, 상세한 분석이 필요한 단백질의 우선순위를 제공할 수 있다는 점에서 의미를 갖는다.

6. 결 론

본 논문에서는 단백질체학 연구의 목적인 임의의 조직에서 어떤 단백질이 발견되는지, 특정 조건 하에서 단백질이 어떻게 상호 작용 하는 지를 분석하기 위해서, 질병에 의해 발현량이 변하는 단백질과 임상 정보와의 연관성 등을 분석하기 위한 데이터 큐브의 척도 및 Discovery-driven 탐색 방법의 응용 방안 등을 제시하였다. 데이터 큐브의 척도와 예외 지표의 값은 기존에 질병 의존 단백질을 검증하기 위해 사용하는 통계적 방법의 문제점을 해결할 수 있는 값으로 제안하였으며, 그 적합성은 통계적 방법과의 비교 실험을 통해 제시하였다. 또한, 질병 의존 단백질의 분석 뿐 아니라 데이터 큐브의 임상을 나타내는 차원을 통해 임의의 임상 정보에 따른 질병 의존 단백질의 변화를 탐색함에 의해, 보다 추가적인 의미 분석도 가능함을 보였다. 그리고 효과적인 큐브

탐색을 지원하기 위해 *SelfExp*과 *InExp*, *PathExp*을 이용하며, *InExp*의 계산량 줄이기 위한 방법을 제안하고, 실제 폐암 데이터로부터 현재 셀의 척도와 *SelfExp*, 임상 분포를 이용한 예외 탐사 방법의 유용성을 실험으로 제시하였다. 실제 실험에 사용된 폐암 데이터는 보통 단백질체학 실험에 쓰이는 환자 샘플 보다는 수가 많았으나 신뢰성 있는 결과를 얻기에는 부족함이 있으며, 데이터가 충분히 확보되면 더 의미 있는 분석이 가능할 것으로 보인다.

참 고 문 헌

[1] S. Y. Cho, K.-S. Park, J.E.Shim, M.-S.Kwon, K.H.Joo, W.S. Lee, J.Chang, H.Kim, H.C.Chung, H.O.Kim, Y.-K.Paik, An integrated proteome database for two-dimensional electrophoreses data analysis and laboratory information management system, Proteomics, 2, 1104-1113, 2002.

[2] Gygi, SP, Rist, B., Gerber, SA, Turecek, F., Gelb, MH and Aebersold, R., Quantitative Analysis of Complex Protein Mixtures Using Isotope-Coded Affinity Tags, Nat.Biotech, Vol.17, No.10, pp.994-9, 1999.

[3] Cagney, G. and Emili, A., De Novo Peptide Sequencing and Quantitative Profiling of Complex Protein Mixtures Using Mass-Coded Abundance Tagging, Nat Biotech., Vol.20, No. 2, 163-70, 2002.

[4] Celis J. E., Rasmussen H. H., Gromov P., Olsen E., Madsen P., Leffers H., Honoré B., Dejgaard K., Vorum H., Christensen D. B., Østergaard M., Haunsø A., Aagaard Jensen N., Celis A., Basse B., Lauridsen J. B., Ratz G. P., Andersen A. H., Walbum E., Kjærgaard I., Andersen I., Puype M., Van Damme J., Vandekerckhove J., The human keratinocyte two-dimensional protein database (update 1995): mapping components of signal transduction pathways, Electrophoresis, 16, 2177-2240, 1995.

[5] Rabilloud, T., Two-dimensional gel electrophoresis in proteomics: Old, old fashioned, but it still climbs up the mountains., Proteomics, 2, 3-10, 2002.

[6] K..S..Park, Y..K..Jeon, S..Y..Cho, D. B..Kim, W..S..Lee, M.-S. Kwon, H. Kim, E. S. Yu, Gao V., Patterson D., B.-D. Han, Y.-K.Paik, Composite Analyses of Metabolic Profiles of Proteins That are Differentially Expressed in Hepatocellular Carcinoma, HUPO-The Second Congress of Human Proteome Organization, Montreal, Canada, 2003.

[7] S. O. Lim, S.-J. Park, W. Kim, S. G. Park, H.-J. Kim, Y. I. Kim, T.-S. Sohn, J.-H. Noh, G. Jung, Proteome Analysis of Hepatocellular Carcinoma, Biochemical and Biophysical Research Communications 291(4), 1031-1037, 2002.

[8] Boer J.M., Huber W.K., Sultmann H., Wilmer F., von Heydebreck A., Haas S., Korn B., Gunawa B., Vente A., Fuzesi L., Vingron M., Poustka A., Identification and classification of differentially expressed genes in renal cell carcinoma by expression profiling on a global human 31, 500-element cDNA array, Genome Research, 11(11), 1161-1170, 2001.

[9] Arnott D., O'Connell K.L., King K.L., Stults J.T., An Integrated Approach to Proteome Analysis: Identification of Protein Associated with Cardiac Hypertrophy, Analytical Biochemistry, 258, 1-18, 1998.

[10] http://www.genebio.com/products/proteome_imaging.html

[11] http://www.nonlinear.com/products/progenesis/samespots/overview.asp

[12] Jane M.C.Oh, Brichory F., Puravs E., Kuick P., Wood C., Rouillard J.M., Tra J., Kandia S., Beer D., Hanash S., A database of protein expression in lung cancer, Proteomics, 1, 1303-1319, 2001.

[13] Agrawal R., Gupta A., Sarawagi S., Modeling multidimensional databases, In Proc. of the 13th Int. Conference on Data Engineering, Birmingham, U.K., 1997.

[14] Sarawagi S., Agrawal R., Megiddo N., Discovery-driven Exploration of OLAP Data Cubes, Research Report RJ 10102(91918), IBM Almaden Research Center, January 1998.

[15] Gray J., Chaudhuri S., Bosworth A., Layman A., Reichart D., Venkatrao M., Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals, Data Mining and Knowledge Discovery, 1, 29-53, 1997.

[16] Jiawei Han and Micheline Kamber, DataMining: Concepts and Techniques, Morgan Kaufmann Publishers, 2000.

심 정 은

e-mail : jjuggeuni@database.yonsei.ac.kr

2001년 인천대학교 전자계산학과(공학사)

2003년 연세대학교 컴퓨터과학과(공학석사)

2003년~현 재 연세대학교 컴퓨터과학과 박사과정

관심분야 : 생물정보학, 데이터웨어하우스, 데이터마이닝

이 원 석

e-mail : leewo@database.yonsei.ac.kr

1985년 미국 보스턴대학교 컴퓨터공학과 (공학사)

1987년 미국 퍼듀대학교 컴퓨터공학과 (공학석사)

1990년 미국 퍼듀대학교 컴퓨터공학과 (공학박사)

1990년~1992년 삼성전자 선임연구원

1993년~1999년 연세대학교 컴퓨터과학과 조교수

1999년~2004년 연세대학교 컴퓨터과학과 부교수

2004년~현 재 연세대학교 컴퓨터과학과 교수

관심분야 : 분산데이터베이스, 미디에이터시스템, 데이터마이닝, 데이터스트림