# 질병 의존 단백질 도출을 위한 데이터 큐브의 응용

김단비[O] 이원석
연세대학교 컴퓨터과학과
{danbi_kim[O], leewo}@amadeus.yonsei.ac.kr

# Application of Data Cube to Identify Differentially Expressed Proteins by Disease

Kim Dan Bi[0] Lee Won Suk
Dept. of Computer Science, Yonsei University

**요 약**

주어진 셀이나 조직에 발현된 단백질 프로파일의 구조적인 분석을 다루는 단백질체학(Proteomics) 연구에 있어서, 질병에 대한 마커 단백질(marker proteins)을 도출(identification)하는 것은 핵심 논점 중 하나이다. 수십 개의 샘플로부터 추출한 셀이나 조직 내에는 수많은 단백질이 포함되어 있으며, 존재하는 단백질의 질병에 의한 발현량(expression level) 변화 및 임상 특성에 의한 영향을 분석하기 위해서 데이터베이스와 데이터 마이닝 기술의 활용이 효과적이다. 본 논문에서는 질병 및 임상 특성에 따른 단백질의 발현량 변화를 분석하기 위한 OLAP 데이터 큐브(Data cube)의 응용 방법과 단백질 데이터의 분석에 적합한 척도(measure)를 제안하고, 유효성을 보인다.

## 1. 서 론

단백질체학 연구는 주어진 셀이나 조직에 발현된 단백질 프로파일의 구조적인 분석을 다루며, 유전자 명령으로 만들어진 단백질체를 대상으로 특정 조건 하에서 단백질의 기능이상 및 구조 변형 유무를 규명하고 질병 과정을 추적하는 것이 목표이다. 따라서, 질병에 따른 마커 단백질을 검증하는 것이 단백질체학의 주요 논점의 하나이며, 질병 의존 단백질(differentially expressed proteins by disease)이 잠재적인 마커 단백질일 수 있다. 이미 간암이나 폐암과 같은 다양한 질병의 진단을 위한 몇몇 마커 단백질들이 제시되었으나, 질병의 진단과 치료를 위한 마커 단백질의 전체 집합을 얻기 위해서는 많은 연구가 필요하다[1].

일반적으로 단백질의 발현량 변화 분석에는 Mann-Whitney test나 Wilcoxon nonparametric paired t-test등의 통계적 방법이 많이 사용되고 있으며[2,3], Melanie IV, Progenesis, PDQuest와 같은 단백질 상용 이미지 분석 소프트웨어에서 제공하는 그래픽을 이용한 방법도 있다. 그러나, 이 방법들은 많은 조직 내에 존재하는 수많은 단백질을 일일이 비교, 분석해야 하는 어려움이 있다. 최근에는 이러한 문제를 극복하기 위해 데이터베이스와 데이터 마이닝 기술을 적용한 다양한 시도들이 진행되고 있다[4].

본 논문에서는 질병뿐 아니라 임의의 임상적 특성에 따라 발현량이 변하는 단백질을 검증하기 위한 OLAP 데이터 큐브의 응용 방법과 단백질의 발현량 변화 분석에 적합한 데이터 큐브의 척도를 제안한다. 또한, 통계적 방법과의 비교 실험을 통해 본 논문에서 제안한 척도가 단백질의 발현량 변화 분석에 적합하며 보다 효과적임을 보인다.

## 2. 임상 특성을 고려한 데이터 큐브

기존의 단백질 발현량에 대한 변화 분석 방법은 질병에 대해서만 적용되었으나, 더 정확한 마커 단백질 도출을 위해서는 질병뿐 아니라 임의의 임상 특성에 대한 변화 분석도 요구된다.

데이터 큐브는 데이터를 다차원 모델로 표현하기 위한 방법으로, 데이터는 분석 관점이 되는 차원(dimension)과 분석 대상이 되는 척도(measure, 또는 fact)로 구성된다[5]. 각 차원은 데이터 큐브의 축을 이루며 계층 구조를 갖는데, 단백질의 발현량 변화 분석을 위해 개개의 단백질을 나타내는 대표 값이 하나의 축이 되며, 각 임상 특성이 다른 축이 된다. 척도는 각 차원을 구성하는 항목들의 조합에 해당하는 데이터들을 나타내는 값이며, 여기서는 질병 의존 단백질 도출을 위한 척도가 된다.

그림 1은 임상 특성에 따른 단백질의 발현량 변화 분석을 위한 데이터 큐브의 예를 나타낸 것으로, 단백질을 나타내는 값 외에, 성별, 나이, 흡연 경력 등의 임상 정보를 이용했으며, 분석 대상 질병의 특성에 따라 다른 차원 구성이 가능하다.

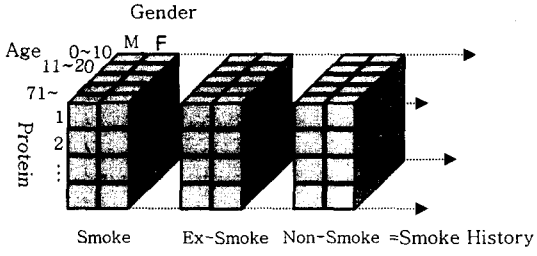


그림 1. 데이터 큐브의 예

### 3. 단백질의 발현량 분석 방법

본 장에서는 단백질 발현량 분석에 대표적으로 사용되는 통계적 분석 방법에 대해 기술하고, 이러한 기존 분석 방법의 문제점을 해결하는 척도를 제안한다.

#### 3.1 통계적 분석 방법: Wilcoxon nonparametric paired t-test

통계적 분석 방법은 비교 대상이 되는 두 집단 간에 차이가 존재할 확률을 구하는 것이다. 여러 샘플에서 추출한 다수의 단백질 중에서 정상 조직과 비정상 조직 간에 발현량 변화가 있는 단백질을 찾는 통계적 방법에는 다음의 두 가지가 있다. 샘플 전체에 대해서, 정상 조직들 내의 임의의 단백질 집합과 비정상 조직들 내의 해당 단백질 집합의 발현량 변화를 비교하는 방법과, 한 샘플에서 추출된 정상 조직 내의 임의의 단백질과 비정상 조직 내의 해당 단백질을 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[2]. 그러나, 단백질 데이터의 특성 상 샘플 간의 변이(variation)가 크므로 이 방법은 적합하지 않다.

후자는 한 샘플에서 추출된 정상과 비정상 조직 내의 단백질을 서로 매치(match)하고, 쌍을 이룬 각 샘플을 모두 매치함에 의해 생성된 두 집단을 비교하는 것으로, Paired t-test방법이 있다. 여기서는 실험에 쓰이는 단백질의 발현량에 대한 데이터가 정규 분포를 이루지 않으므로, Wilcoxon nonparametric paired t-test방법이 적합하다. 이 방법은 쌍을 이룬 단백질의 발현량 차이의 크기에 따라 증감 별로 순위를 취하여 두 집단 간의 변화를 비교하는 것으로, 차이가 있을 가능성을 결과값(*p-value*)으로 한다. 여기서 가설은 두 집단 간에 차이가 존재하지 않는 것이고, 결과값 $P=0.05$이하이면 95%수준에서 가설을 기각한다고 판단한다[3].

#### 3.2 U_G_O_ratio

앞 절에 설명한 통계적 방법에는 다음의 문제점이 있다. 정상과 비정상 조직 내의 단백질 발현량의 비교에 쓰이는 기준이 두 값 간의 차이이므로, 실험 상의 외부 변화 요소에 의한 단백질의 발현량과 관계 없는 미약한 차이도 발현량의 변화로 취급하게 된다. 그러나, 미약한 차이를 배제하기 위하여 해당 샘플을 필터링하는 것은 전체 데이터의 특성을 잃게 되므로 부적합하다. 또한, 샘플 간의 변이를 고려하지 않고 차이의 크기로 순위를 취하므로, 다수의 샘플에서 감소하는 경향을 보이는 단백질 발현량이 소수의 샘플에서 큰 차이로 증가하는 경우, 두 경향이 상쇄됨에 의해 이 단백질은 다르게 발현되는 단백질로 판단되지 않는다.

일반적으로 단백질 발현량의 차이가 25%이내인 경우는 생물학적으로 중요하지 않다고 판단한다. 또는, 실험 대상이 되는 질병의 종류나 샘플의 변이 및 실험자의 판단에 따라 보다 강력한 임계값(threshold)을 사용하기도 한다. 동일한 단백질 $p_n$과 $p_t$에 대해서, $p_n$이 정상 조직 내의 단백질을, $p_t$가 비정상 조직 내의 단백질을 나타내고, *ExpLevel(p)*이 $p$의 발현량을 나타내는 값이라고 할 때, 발현량 차이의 임계값 $x$에 따라 단백질 발현량의 변화는 다음과 같이 분류한다. 정상에 비해 비정상의 단백질 발현량이 $x$배 이상으로 감소한 경우는 Under-expressed protein으로, $x$배 이상으로 증가한 경우는 Over-expressed protein으로, $x$배 이내로 변화한 경우는 General protein으로 분류한다.

$$\begin{cases} \text{Under-expressed protein, if } \dfrac{ExpLevel(p_t)}{ExpLevel(p_n)} \le \dfrac{1}{x}. \\ \text{General protein, if } \dfrac{1}{x} < \dfrac{ExpLevel(p_t)}{ExpLevel(p_n)} < x. \\ \text{Over-expressed protein, if } \dfrac{ExpLevel(p_t)}{ExpLevel(p_n)} \ge x. \end{cases}$$

주어진 여러 샘플에서 동일한 단백질이 어떠한 경향으로 발현량이 변화하는지 분석하기 위한 척도는 다음과 같이 정의한다.

$$U\_G\_O\_ratio(p) = \max(O\_ratio(p), U\_ratio(p))$$

여기서, 임의의 단백질 $p$가 Under-expressed protein인 샘플의 수를 *U_cnt(p)*, General protein인 샘플의 수를 *G_cnt(p)*, Over-expressed protein인 샘플의 수를 *O_cnt(p)*라 할 때,

$$O\_ratio(p) = \frac{O\_cnt(p)}{O\_cnt(p)+G\_cnt(p)+U\_cnt(p)}$$ 이고,

$$U\_ratio(p) = \frac{U\_cnt(p)}{O\_cnt(p)+G\_cnt(p)+U\_cnt(p)}$$ 이다.

*U_G_O_ratio*의 값이 0.5이상인 경우, 임의의 단백질 $p$는 전체 샘플의 50%이상에서 Under-expressed되거나 Over-expressed된 것이며, 따라서 해당 단백질은 발현량 변화를 보인다고 판단한다. *U_G_O_ratio*로부터 다르게 발현하는 단백질을 판단하기 위한 임계값은 0.5이상의 값으로 실험자의 선택에 따라 결정할 수 있다.

### 3. 실험 및 결과

실험에서는 간암에 관련된 환자 51명의 데이터와 폐암에 관련된 환자 20명의 데이터를 각각 이용했다. 환자의 질병 관련 조직에서 얻은 정상 조직과 비정상 조직으로부터 2-DE방법과 2-DE 이미지 소프트웨어를 이용하여 실험 데이터를 추출하였는데, 본 실험에서는 이미지 소프트웨어로써 간암 샘플에 대해서는 Melanie Ⅲ를, 폐암 샘플에 대해서는 Progenesis를 이용했다. 여기서, 단백질의 발현량을 나타내는 값은 %Vol 값을 이용하였다.

그림 2는 51개의 간암 데이터 샘플 중10개 이상의 샘플에서 쌍으로 나타난 단백질에 대한 데이터 큐브의 결과 일부를 나타낸 것이다. 큐브의 셀에는 발현량 차이의 임계값 $x=1.5$로 했을 때의 *U_cnt(p)*, *G_cnt(p)*, *O_cnt(p)*를 나타내었으며, *U_G_O_ratio*의 값을 척도로 하여 *U_G_O_ratio*가 0.5이상인 것을 발현량이 변한다고 판단하였다. Under-expressed protein이라고 판단된 것은 굵은 글씨로, Over-expressed protein이라고 판단된 것은 음영으로 나타내었다.

그림에서, 1-D데이터 큐브에서 의미를 보이지 않던 539번 단백질이 남성인 경우 Under-expressed로 판단된 것을 볼 수 있다. 이와 같이, 임상 특성을 차원으로 하는 데이터 큐브를 사용함으로써, 임상 특성의 고려 없이는 발현량이 증가하기도 하고 감소하기도 하여 별 의미가 없다고 판단되는 단백질인 경우

도, 성별로 drill down했을 때 남성인 경우는 발현량이 증가하고 여성인 경우는 감소하는 등의 의미 있는 분석이 가능하다.

| Protein | U_G_O_Cr |
|---|---|
| 407 | 7_4_1 |
| 538 | |
| 539 | 7_5_3 |
| 542 | 14_5_5 |
| 717 | 12_5_4 |
| 877 | |
| 883 | 4_5_5 |
| 886 | 4_4_4 |
| 1198 | 5_5_2 |
| 1200 | 10_8_9 |

⇒

| Protein | Normal | Disease |
|---|---|---|
| 407 | 4_2_1 | 3_1_0 |
| 538 | | 2_5_5 |
| 539 | 7_1_2 | 0_4_1 |
| 542 | 10_2_3 | 4_3_2 |
| 717 | 8_3_4 | 4_1_0 |
| 877 | 1_6_4 | |
| 883 | | 3_2_1 |
| 886 | 4_3_4 | 0_1_0 |
| 1198 | 4_2_2 | 1_1_0 |
| 1200 | 5_7_6 | 5_1_3 |

1-D 데이터 큐브 ⇒ 2-D 데이터 큐브

그림 2. 데이터 큐브 결과

그림 3은 20개의 폐암 데이터 샘플 중 7개 이상의 샘플에서 쌍으로 나타난 단백질에 대해서, 기존의 통계적 방법과 제안한 *U_G_O_ratio*에 따른 발현량이 변하는 단백질에 대한 분포를 나타낸다. (Stat- U_G_O)는 통계적 방법의 95%신뢰 수준에서만 발현량이 변한다고 판단되는 단백질을, (U_G_O-Stat)는 발현량 차이의 임계값 $x$가 2.0이고 0.5이상의 *U_G_O_ratio*에서만 발현량이 변한다고 판단되는 단백질을 나타낸 것이다. (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.

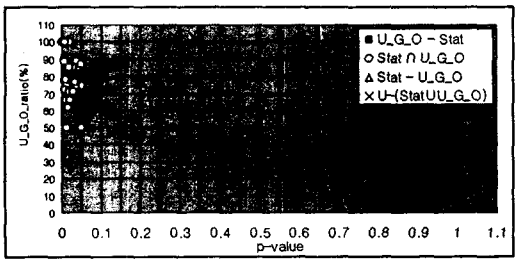


그림 3. 단백질의 결과 분포 비교

그림 4은 통계적 방법에서 단백질의 발현량으로 보기 힘든 미약한 발현량 차이가 변화로 인식된 단점을 나타낸 그래프이다. 발현량 차이의 임계값 $x$에 대해 전체 protein에서 General protein으로 분류되는 비율의 평균치를 나타낸 것으로, (Stat-U_G_O)의 General protein의 비율이 다른 데이터에 비해 크며, $x$값이 증가함에 따라 비율이 증가함을 알 수 있다.

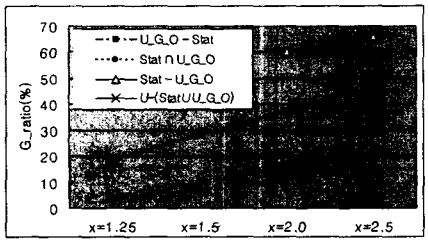


그림 4. 통계적 방법의 미약한 발현량 변화 인식 문제

그림 5는 (U_G_O-Stat)에 해당하는 데이터가 통계적 방법에서는 의미 있는 데이터로 판단되지 않는 이유를 나타낸 것이다. 그래프의 y축은 하나의 단백질이 샘플의 특성에 따라 발현량이 증가하기도 하고 감소하기도 하는 경우, 발현량 증감의 상쇄 정도를 나타낸 것으로, 다른 데이터보다 (U_G_O-Stat)와 (U-(Stat∪U_G_O))에 해당하는 데이터의 상쇄 정도가 큼을 알 수 있다. 즉, 통계 방법에 따른 척도를 이용하는 경우, 소수의 경향에 의해 다수의 경향이 상쇄되는 단점이 있다.

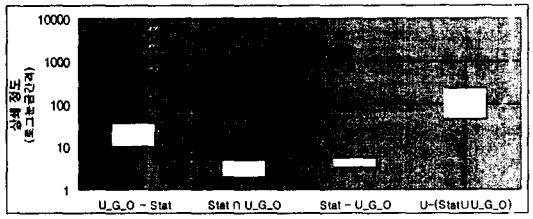


그림 5. 통계적 방법의 상쇄 효과

위의 두 실험에 의해서 통계적 방법의 단점을 확인할 수 있으며, 제안한 척도가 이러한 단점을 해결함에 의해 보다 적합함을 알 수 있다.

## 5. 결 론

본 논문에서는 질병에 따라 다르게 발현되는 단백질을 검증하기 위해 기존에 사용하는 통계적 방법의 문제점을 해결할 수 있는 척도를 제안하고, 이 척도를 이용한 데이터 큐브의 응용 방안을 보였다. 제안한 척도는 실험자의 판단에 따라 데이터에 적합한 임계값을 설정할 수 있으며, 척도의 값에 따라 실제 단백질 검증 실험을 위한 우선순위 지정이 가능하다. 또한 데이터 큐브를 이용함으로써, 질병뿐 아니라 임상 특성에 따른 분석을 통해 단백질과 임상 특성에 따른 추가적인 의미 분석 효과도 기대할 수 있다.

## 참고 문헌

[1]K.S Park, Y.K Jeon, S.Y. Cho, D.B. Kim, W.S. Lee, Y.K Paik, 밃omposite Analyses of Metabolic Profiles of Proteins That are Differentially Expressed in Hepatocellular Carcinoma□ HUPO, The second Congress of Human Proteome Organization, Sep. 2003.

[2]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, 밃roteome Analysis of Hepatocellular Carcinoma□ Biochemical and Biophysical Research Communications 291, 1031-1037, 2002.

[3]David Arnott, Kathy L. O뮰onnel, Kathleen L.King, John T.Stults, 밃n Integrated Approach to Proteome Analysis: Identification of Protein Associated with Cardiac Hypertrophy□ Analytical Biochemistry 258, 1-18, 1998.

[4]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.Pa:k, 밃n integrated proteome database for two-dimensional electrophoreses data analysis and laboratory information management system□ Proteomics, 2, 1104-1113, 2002.

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