DBMS 기반 식물 돌연변이 단백질체 분석을 위한 시스템 설계 및 구현

허정호°, 박춘구, 박윤주, 최지인, 정동수, 남홍길
포항공과대학교 생물학전문연구정보센터
(hjh, madreach, oct1001, jiin, viroid97, hgn)@bric.postech.ac.kr

Design and Implementation of DBMS-based Plant Mutant Proteome Analysis System

Jeong Ho Huh°, Chun Gu Park, Yoon Ju Park, Ji In Choi, Dong Su Jeong, Hong Gil Nam
Postech , BRIC

요 약

생물학의 주된 목적이 세포 안에서의 유전자의 기능을 알아내는 것이므로 유전자의 실제 기능을 담당하는 단백질에 대한 연구는 필연적인 것이다. 근래에 이르러 2DE 등의 실험방법 개선 등으로 대량의 단백질에 대한 연구가 실현화 됨에 따라, 프로트오믹스(proteomics)의 연구가 활성화되었으며, 본 실험팀 역시 식물 애기장대의 돌연변이들을 이용한 프로티오믹스 연구를 진행중이다. 프로티오믹스 연구는 실험 특성상 대량의 단백질 정보들을 얻게 되므로 데이터의 수작업 분석이 불가능하다. 그리하여 우리는 식물 애기장대(arabidopsis)의 특성과 2DE 실험에서 파생되는 여러 데이터들을 반영하여 정보를 효율적으로 관리할 수 있는 데이터베이스를 설계 구축하였으며, 나아가 각 표현형을 대표할 수 있는 표지 단백질을 선별해내는 분석모듈을 구현하였다.

1. 서론

분자생물학의 발전과 분석기기들의 자동화로 인해서 인간을 포함한 다양한 종들의 게놈(genome- 유전자 집합체) 프로젝트들이 이미 완료되었고, 또 진행과정 중에 있다. 이렇게 밝혀낸 염기서열을 분석하여 의미 있는 유전자를 밝혀 내는 연구가 활발히 진행 중이다[1]. 이러한 연구 분야를 "Functional Genomics"(유전체 기능 분석학)라고 한다.

지금까지 연구는 대량의 DNA의 분석이 주가 되었으나, 요즘 들어 실제 기능을 구현하는 단백질(protein)들에 대한 연구 또한 주목 받고 있다. 유전 정보는 DNA가 가지고 있지만, 그로 인해 발현되는 단백질들이 세포 내에서 언제, 어떻게, 어느 정도의 양으로 기능 하는지는 DNA 정보로 알 수 없기 때문이다. 체내에 있는 세포들이 가지고 있는 게놈은 모두 같지만, 세포마다 프로티옴(proteome - 단백질 집합체)은 다르다. 또한 같은 종류의 항상 같은 프로티옴을 보이는 것이 아니라 때와 환경에 따라 다른 프로티옴을 보인다. 즉, 게놈은 정적인 개념이라면 프로티옴은 동적인 개념인 것이다[2]. 그렇기 때문에, 기능을 연구하는 목적으로는 단백질을 연구하지 않을 수 없다.

그러나 최근까지 프로티옴의 연구가 활발하지 못했던 이유는, 세포 내의 단백질을 하나 둘이 아닌, 전체적으로 관찰할 수 있는 실험 방법이 여의치 않았기 때문이다. 그러나 최근 들어 2DE 등의 실험 방법이 발달하고 체계화됨에 따라 프로티옴의 연구가 활발하게 되었다.

본 실험팀은 식물의 한 종인 애기장대의 다양한 돌연변이들의 표현형과 그 돌연변이의 프로티옴을 밝혀내어, 특정 표현형을 나타내는 유전자들의 기능을 유추, 발굴해 내는 연구를 진행 중에 있다. 애기장대는 식물의 대표적인 시료식물로서 전세계의 많은 연구팀이 이를 대상으로 연구하고 있다[3]. 이 논문은 이런 계획의 일환으로 proteome연구의 많은 데이터를 저장하고, 특정 표현형의 표지로 쓰일 수 있는 단백질 (표지단백질 - marker protein)을 식별하는 기능을 데이터 베이스 시스템을 이용하여 설계, 구현하는 것에 목적을 두고 있다.

2장에서 프로티옴 연구의 기본이 되는 2DE(2D gel Electrophoresis)와 단백질 동정, Staining 등의 방법에 대해 알아보고, 3장에서 시스템의 설계와 구현을 설명, 4장에서 결론과 향후 연구에 대해 논한다.

2. 관련연구

2.1 2DE(2D gel Electrophoresis)

2D gel electrophoresis는 현재 수천개의 단백질을 동시에 분리해낼 수 있는 유일한 방법이다. 단백질은 각각의 고유한 전하(charge)와 질량(mass)을 갖고 있으며, 2D는 그것을 이용하여 분류한다. pH gradient에서 단백질의 전하(charge)를 이용하여 1차로 분류하고, 질량을 이용하여 2차로 분류한다. 그래서 2D map이라 불리는 sample을 얻는다. 이 곳의 Spot은 하나의 단백질을 뜻한다. 2DE 실험 이후, 특정 spot을 적당하게 잘라내서 단백질 동정의 작업

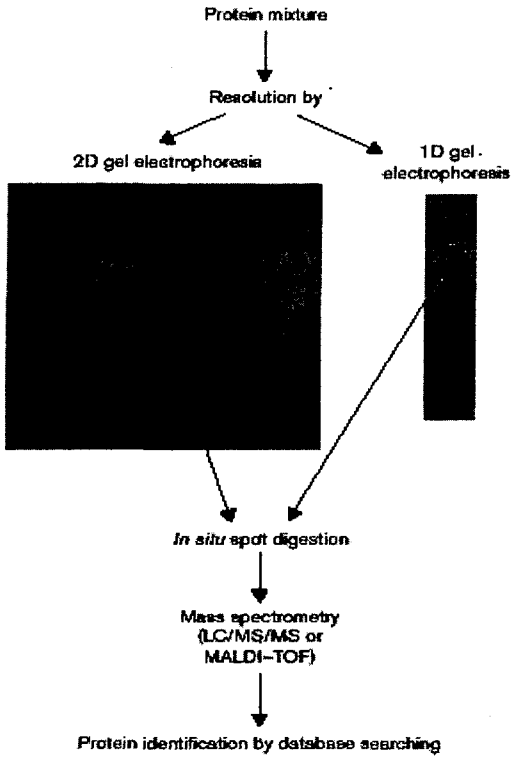

그림1. Proteome 실험의 전체적인 과정[2]

을 한다. 동정 작업은 2DE에 비해 시간이 오래 걸리므로 프로티옴 실험의 병목 작업이라 할 수 있다[4].

2.2 Protein identification

단백질 동정 작업에는 다양한 방법들이 쓰인다. 최근에는 단백질 단편들의 정확한 질량을 재서 단백질의 정체를 알아내는 MS(Mass Spectrometry) 방법이 빠르고 정확한 방법으로 각광 받고 있는 추세이긴 하지만, 여전히 다른 방법들도 많이 쓰이는 편이다. MS 이외에는 단백질의 20개 아미노산의 구성 성분을 분석하는 방법과 Edman degradation등을 이용하여 단백질 말단 부분의 단편적인 서열을 알아내는 방법이 있다. 그러나 대부분의 경우 정확성을 더욱 높이기 위해 하나만의 방법을 고집하지 않고, 여러 개의 방법을 동시 적용하는 때가 많다. 그림 1은 2DE부터 단백질 동정까지 간략적인 실험의 단계를 표현한 것이다.

2.3 Protein quantitation

단백질 정량에는 주로 염색 방법(staining)이 쓰인다. 주로 쓰이는 것이 silver staining과 Coomassie Blue staining이 있다. Silver staining은 단백질 양이 0.04ng/mm2 ~ 2ng/mm2 사이에 있을 경우 염색 정도가 단백질 양과 선형 비례한다. 반면, Coomassie Blue staining은 10-200ng일 경우에 선형 비례한다. 그러므로 실험에 쓰인 샘플의 양을 생각해서 염색 방법을 적절히 사용한다[4].

3. 시스템의 설계 및 구현

3.1 시스템 개요

애기장대 프로티옴 실험의 전체적인 흐름은 그림 2와 같다. 이 그림에서 박스에 둘러싸인 부분이 Proteome DB system에 해당하는 부분이다.

실험을 통해 얻어진 단백질들의 동정 정보와 단백질량, 2D map에서의 위치 등에 대한 종합적인 정보들은 하나의 profile로 만들어진다. 이 profile들은 항목에 따라 구분되어 DB에 저장된다. 이 정보들은 표현형의 표지 단백질을 구분해내기 위한 데이터로 쓰이게 되고 시스템은 주어진 기준에 따라 표지 단백질들을 분류해 낸다. 실험이 계속 진행되면서 새로운 정보가 들어올 때마다, 새로 표지 단백질을 찾아내는 작업을 계속하게 된다.

이렇게 얻어진 표지 단백질에 대한 정보는 새로운 식물 돌연변이(mutant)의 프로티옴을 연구할 때 비교 정보로 쓰인다. Random profile은 아직 어떤 정보도 알려지지 않은 식물 돌연변이에 대한 프로티옴 정보를 뜻한다. 시스템은 표지 단백질을 이용해 random profile을 자동적으로 분석하여 그 결과를 사용자에게 알려준다. 사용자는 그 정보를 통해 새로운 사실을 알게 되거나 그를 이용하여 실험 계획을 보강, 수정할 수 있다.

그림을 통해 알 수 있듯이 시스템의 세부 구성은 실험 데이터를 저장할 데이터베이스, 그 정보를 통해 표지 단백질을 선별하는 평가 모듈, 그리고 선별된 표지 단백질을 통해서 random profile을 분석하는 모듈로 나눌 수 있다.

3.2 데이터베이스 구상

데이터베이스는 시스템에 가장 기본이고 주된 것이라 할 수 있다. 그림 3은 시스템의 근간인 데이터베이스의 E-R diagram이다.

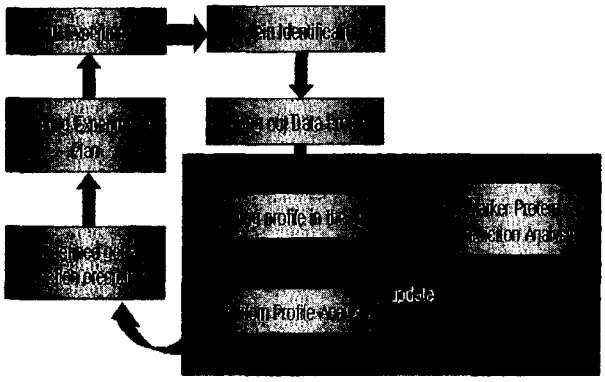

그림2. 애기장대 프로티옴 실험의 정보 흐름도

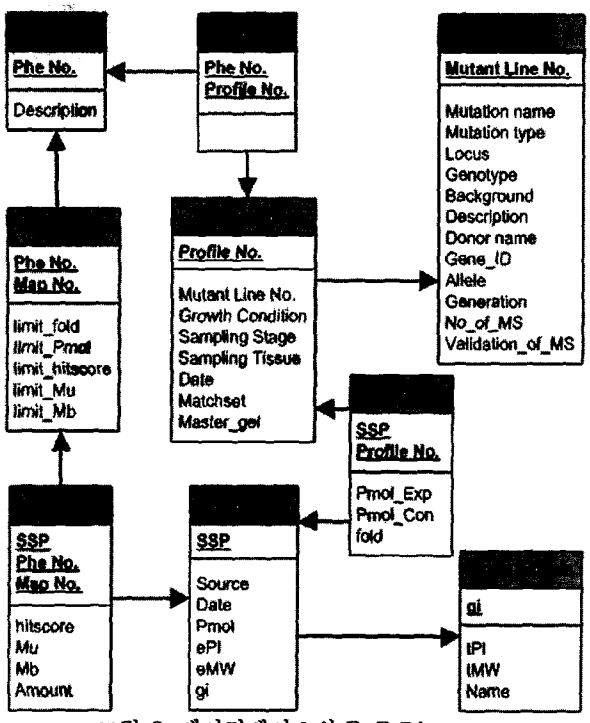

그림 3. 데이터베이스의 E-R Diagram

실험에 쓰인 돌연변이 식물에 대한 정보는 Mutant가 갖게 되고, 그에 대한 실험의 일반적인 정보는 Profile_Info와 Protein_List 테이블에 저장되게 된다. 각 실험에 쓰인 식물들은 각각 독특한 표현형을 갖고 있으며, 그에 대한 정보는 Expression에 저장된다. Phenotype 테이블은 실험과정에서 나올 수 있는 표현형의 전체 집합이라 할 수 있다. Marker_Map은 평가 모듈에 적용될 표지 단백질 선정 기준을 저장한 곳이다. 평가 모듈을 통해서 얻은 각 표현형의 표지 단백질에 대한 정보는 Marker_List에 저장되게 된다. Master_Map은 실험과정에서 나올 수 있는 모든 단백질에 대한 전체 집합이라 할 수 있으며, Public_Protein은 단백질과 관련된 다른 데이터베이스의 정보를 연결한 것이다.

3.3 시스템 구현

본 시스템은 다음과 같은 환경에서 구현되었다.

- ♦DBMS: MySQL 3.23
- ♦OS: Wow Linux 7.0
- ♦Programming Language: JDK 1.4.0

JDBC를 이용해 데이터베이스를 연결하고 JSP를 사용해 어플리케이션을 구현하였다. 사용자는 웹을 통해 DB를 이용하고 분석툴을 사용하게 된다.

4. 결론 및 향후 연구

프로티옴 실험의 데이터는 그 특성상 대량의 정보를 얻게 되므로 이 데이터를 저장, 분석하기 위한 데이터베이스가 필요하다. 우리는 실험에서 파생되는 모든 종류의 데이터를 담기 위한 관계형 데이터베이스를 구축하였다. 또한 아직까지 연구되지 않은 돌연변이에 대한 실험 데이터 분석을 위해, 표지 단백질을 선별하기 위한 분석모듈을 추가하였다.

파일 시스템이 아닌 데이터베이스를 구축한 주된 이유 중 하나는 향후 진행될 데이터마이닝 단계를 위한 것이다. 본 실험팀은 데이터마이닝의 한 분야인 classfication을 적용한 모듈을 추가할 예정이다[5]. 아직 알려지지 않은 기능의 유전자를 이를 통해 어느 정도 분류해낼 수 있을 거라 기대한다.

또한 모든 표현형에 대해서 각각 단백질들의 고유한 패턴을 나타낸 뒤에 시계열 매칭(time series matching)알고리즘을 이용해 비교하는 방법을 연구 중에 있다[6]. 이를 통해 단백질의 다양한 기능 정보를 어느 정도 예측할 수 있으리라 본다.

5. 참고 문헌

[1] 백용기, *프로테오믹스의 연구방향과 국내 프로테오믹스 연구의 활성화*, 한국분자생물학회 뉴스지, vol12,20-29, 2000

[2] 박성구, *프로테오믹스 연구의 필요성과 추진 방향*, Biozine , 2001.2

[3] Birgit Kersten et al, *Large-scale plant proteomics*, Plant Molecular Biology, vol48, 133-141, 2002

[4] M.R. Wilkins et al, *Proteome Research: New Frontiers in Functional Genomics*, Springer, 1997

[5] Amanda Clare et al, *Machine learning of functional class from phenotype data*, Bioinformatics, vol18 no1, 160-166, 2002

[6] 김태훈, *시계열 데이터베이스에서의 모양 n 기반 서브시퀀스 매칭*, 28회 정보과학회