프로테옴 라이브러리의 연합을 통한 이차원 전기영동데이터베이스 구축도구 개발

박진수*, 정채영*, 김종준*, 김재원**, 이창원**, 배종민*
*경상대학교 컴퓨터과학과, **경상대학교 미생물학과
{lanez04, wcdi}@sys.gsnu.ac.kr, sinby816@soback.kornet21.net
{jwkim, cwlee, jmbae}@nongae.gsnu.ac.kr

Development of a Tool for Building a 2-DE Database Based on the Federation of Proteome Libraries

Jin-Soo Park*, Chai-Young Jeong*, Jong-Joon Kim*, Jae-Won Kim**, Chang-Won Lee**, Jong-Min Bae*
*Dept. of Computer Science, **Dept. of Micro Biology. Gyeongsang National University

요 약

인간 게놈프로젝트가 거의 완성되어 감에 따라서 해독된 유전자의 암호를 바탕으로 기능을 밝히는 연구가 많이 진행되고 있다. 그 중에서 디스플레이 프로테오믹스에 관한 연구가 주목을 받으면서, 세계적으로 2-DE 데이터베이스가 많이 구축되었다. 이 데이터베이스를 구축하기 위해서는 이미 개발된 많은 프로테옴 관련 데이터베이스를 참고해야 한다. 따라서 2-DE 데이터베이스를 빠르고 효과적으로 구축하기 위해서는 자동화된 구축도구가 필요한데, 이 도구를 개발하기 위해서는 데이터베이스 통합이 필수적으로 요구된다. 본 논문에서는 생물학적 실험을 통하여 얻어진 펩티드 질량으로써, 다수의 펩티드 질량검색기를 통합하여 자동으로 단백질을 인식하고, 인식된 단백질에 대한 정보를 제공하는 다수의 프로테옴 관련 데이터베이스의 통합을 통해서 2-DE 데이터베이스를 자동으로 구축하는 도구와 데이터베이스 구축결과를 제시한다.

1. 서론

인간 게놈프로젝트가 거의 완성되어 감에 따라서 해독된 유전자의 암호를 바탕으로 기능을 밝히는 연구, 즉, functional genomics에 대한 연구가 활발히 진행되고 있다. functional genomics의 기술적 도구로는 DNA 칩, 단백질 칩, display proteomics 등이 있는데, 각 기술마다 장단점이 있다. 그 중에서 display proteomics는 유전자 기능해석을 위한 강력한 도구로 주목받고 있다. 프로테옴(proteome)이란, PROTEin과 genOME의 합성어로서, 어떤 조직이나 세포에서 발현되는 단백질의 총체를 의미한다. 이것이 주목받는 이유는 생명체가 살아가는데 필요한 일을 직접 수행하는 것은 유전자의 산물인 단백질이므로, 질병을 포함한 모든 생리현상은 결국 단백질에 의해서 지배되기 때문이다.

프로테오믹스 연구방향 중의 하나는 표준 gel양식을 사용해서 2차원전기영동(2-Dimensional Electrophoresis, 2DE)으로 단백질을 분리한 다음, 이를 단백질 분석기술과 접목하여 2-DE gel 상에 나타난 모든 단백질을 게놈정보의 특정 유전자와 연결시켜서 유전자-단백질 대응관계를 데이터베이스로 구축하는 작업이다. 단백질을 분석하고 동정하는 과정에는 세계적으로 많이 개발되어 있는 단백질 관련 데이터베이스를 활용해야 한다.

단백질관련 데이터베이스는 저장되어 있는 정보의 종류에 따라서, 단백질서열 데이터베이스, 패턴/프로필 데이터베이스, 2DPAGE(two-dimensional polyacrylamide gel electrophoresis) 데이터베이스[2], 3-D구조 데이터베이스, post-translational modification 데이터베이스, 게놈 데이터베이스, 메타블릭 데이터베이스 등으로 나뉘어 진다.[4] 그리고 이미 개발된 데이터베이스의 개수를 합치면 수 백 개에 이른다. 이에 따라서 많은 이질의 데이터베이스를 통합하는 문제가 단백질연구분야에서 중요한 이슈중의 하나가 되었다.[4]

본 연구의 궁극적인 목표는 프로테오믹스 기술을 사용하여, 질병상태에서만 특이적으로 나타나는 단백질을 찾아내어, 진단마커로서 사용하거나, 신약개발의 대상으로 삼는 것이다. 이를 위한 하나의 과정으로서 질병관련단백질을 고속으로 탐색 발굴하기 위하여 프로테옴 인식과정이 자동화되어야 하고, 인식된 결과는 데이터베이스에 저장되어야 한다. 본 논문에서는 다수의 펩티드 질량검색기를 통합하여 자동으로 단백질을 인식하고, 펩티드 질량검색을 통해서 식별된 단백질에 대한 정보를 제공하는 다수의 데이터베이스의 통합을 통해서 2-DE 데이터베이스를 자동으로 구축한 결과와 그 개발도구를 제시한다.

2. 관련연구

프로테옴 데이터베이스 통합에 대한 기초적인 모양은 각 데이터베이스에서 상호참조를 데이터베이스 엔트리로 두는 것이다[5]. 이는 세계적으로 가장 유명한 단백질 서열 데이터베이스인 SWISS-PROT에서 광범위하게 사용되고 있다. 이 방법은 비록 많은 제약점을 가지고 있기는 하지만, 다른 탐색도구와 결합될 때 중요한 자산이 될 수 있다. 웹 기반의 대표적인 통합 데이터베이스로는 제네바 대학의 ExPASy(http://www.expasy.ch)를 들 수 있다. 이는 SWISS-PROT, SWISS-2DPAGE, SWISS-3DIMAGE, PROSITE 데이터베이스에 대하여, 하이퍼링크를 기반으로 하여 서로 유기적으로 동작한다. 그리고 이들 데이터베이스에서 제공하는 검색결과에는 다른 데이터베이스에 대한 많은 하이퍼링크를 제공한다. 세계적으로 퍼져있는 2-DE gel 이미지 데이터베이스에 대한 연합(federation)을 제공하는 시스템이 있다.[6] 이 시스템은 2-DE 데이터베이스에 대한 메타검색을 지원할 뿐 아니라, 검색된 2-DE 이미지를 가지고, 사용자의 이미지와 비교할 수 있는 기능도 제공한다.

프로테옴 데이터베이스의 통합기술은 디지털라이브러리의 연합(federation)문제와 유사한데, 연합의 핵심문제로는 정보원의 자동선택, 질의어 생성, 검색결과의 통합과 랭킹 문제가 있다.[3] 프로테옴 데이터베이스의 연합에 있어서는 그 복적에 따라서 위의 세 가지 문제를 모두 해결해야 할 필요가 없을 수 있다. 예를 들어 본 연구에서 개발된 도구에서는 정보원을 미리 정하여 두는 것으로 충분하기

본 연구는 과기부의 21세기 프론티어사업 중 인간유전체 기능연구사업단의 지원으로 수행되었음.

때문에 정보원의 자동선택 문제는 다루지 않는다.

대표적인 2-DE 데이터베이스 구축도구로는 make2ddb[1]가 있다. 이 도구는 SWISS-2DPAGE와 유사한 검색기능을 가진 2-DE 데이터베이스 구축도구인데, 데이터베이스 연합의 기능이 약하고, 사용자의 개입이 많이 필요하다. 본 논문에서는 다수의 펩티드질량검색기를 연합하고, 이를 통해서 인식된 단백질을 바탕으로 자체 구축한 임상병리데이터베이스, 단백질서열데이터베이스, DNA 데이터베이스, 문헌정보데이터베이스를 통합하는 도구를 개발하여 2-DE 데이터베이스를 구축한 결과를 제시한다.

3. 문제 정의

생물학적 샘플에 대하여 2차원전기영동(2-Dimensional Electrophoresis, 2DE)으로 단백질을 분리하면 2-DE gel 상에 그림 1과 같이 점으로 분리된다. 문제는 각 점들은 무슨 단백질인지를 알아내어야 하고, 식별된 단백질에 대해서는 관련된 정보를 관리하고 단백질에 대한 검색서비스를 다양한 방법으로 제공하는 것이다. 그림 1에서 특정 점에 접근했을 때 그 점에 대한 단백질의 이름을 보여주고 있으며, 그 점을 클릭하면 그 점에 대응되는 단백질에 대한 모든 정보를 보여준다. 이와 같은 데이터베이스를 2-DE 데이터베이스라고 한다.

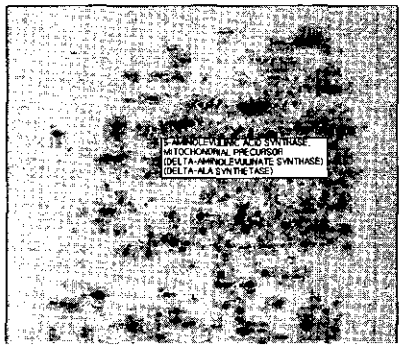

그림 1. 2-DE gel 이미지

본 연구에서 사용된 단백질 분석과정에서, 단백질을 분리하면 점들의 집합으로 구성된 2-DE 이미지가 얻어지고, 이미지 상의 각 점에 대한 펩티드 질량 값의 집합을 실험을 통해서 얻을 수 있다. 그리고 각 단백질에 대한 펩티드 질량을 이론적으로 계산해 둔 데이터베이스가 다수 있는데, 실험을 통하여 얻은 펩티드 질량 값의 집합이 주어졌을 때 그 값을 모든 단백질에 대하여 이론적으로 구해 둔 질량과 비교하여 가장 일치하는 단백질을 결정하여 사용자에게 그 결과를 보여주는 도구도 다수 있다. 이 도구들은 다수의 펩티드 질량 데이터베이스를 지원하고 있으며, 각각 고유의 랭킹 알고리즘을 사용하고 있어서, 추정되는 단백질의 순서가 서로 다르게 나타나는 경우가 많이 있다. 따라서 사용자가 결정하기 어려운 단백질의 경우에, 검색된 결과에 따르는 부수적인 생물학적 정보를 참고하기도 하고, 이들 도구들에게 입력되는 다양한 파라미터를 바꾸어서 다시 검색하기도 하고,(즉 질의어 수정) 검색된 이론적 질량 값을 가진 데이터베이스를 변경시키기도 하며, 각각의 인식도구들을 모두 개별적으로 실험하여 그 결과를 통합함으로써 주어진 질량 값에 대한 단백질을 결정한다. 최악의 경우에는 실험에서 얻은 질량 값 자체에 대한 오류가능성 때문에 이를 다시 얻기 위하여 생물학적 실험을 다시 하기도 한다.

한편, 2-DE 이미지 상의 한 점에 대한 단백질이 결정되면, 그 단백질에 대한 정보를 모으기 위하여, 앞에서 나열한 다양한 프로테옴 관련 데이터베이스에 접근한다. 그리고 얻어진 정보를 바탕으로 자신의 2-DE 데이터베이스를 구축한다. 그리고 마지막으로, 2-DE 이미지에서 주어진 점을 클릭했을 때 그 점에 대한 모든 정보를 제공할 수 있도록 이미지 맵을 만든다. 이러한 검색기능은 점에 대한 클릭 뿐 아니라, 단백질이름, 자체적으로 주어진 고유번호 등에 따라서 검색을 할 수 있도록 한다.

통합의 대상을 살펴보면, 각 단백질에 대한 펩티드 질량을 이론적으로 계산해 둔 데이터베이스들, 이들 데이터베이스를 활용해서 단백질을 인식하는 도구들, 프로테옴 관련 데이터베이스들, DNA 데이터베이스들, 그리고 문헌정보데이터베이스들이다. 위에서 제시한 모든 과정을 자동화 할 수 있는 도구를 개발하기 위해서는 이들 질량검색도구 및 데이터베이스들의 통합과 지역 데이터베이스의 자동구축이 이루어져야 한다. 본 논문에서 소개하는 도구는, 사용자가 2-DE 이미지 상의 한 점을 클릭하고 그 점에 대한 펩티드 질량을 입력하면, 위에서 나열한 과정을 자동으로 수행하여 데이터베이스를 구축하고, 이미지 맵을 통한 검색도 자동으로 이루어지도록 하는 것이다.

4. 시스템 구성

그림 2는 개발된 데이터베이스 구축도구에서 통합한 프로테옴 라이브러리의 구성을 나타낸다. 최초에 단백질 인식도구들 각각에 대하여 스레드를 생성시켜서 분산검색을 시작한다. 그 결과로서 단백질이 인식된 이후에는 나머지 데이터베이스들에 대한 스레드를 생성시켜서 필요한 정보를 수집한다.

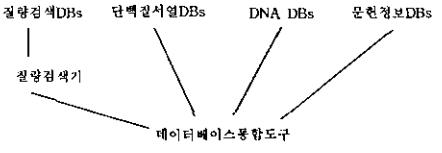

그림 2. 통합대상 프로테옴 데이터베이스

그림 3은 개발된 자동화 도구의 개략적인 내부 구조이다.

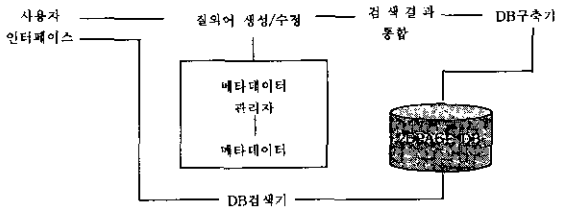

그림 3. 2-DE 데이터베이스 구축도구의 내부 구조

사용자가 2-DE 이미지 상의 한 점을 선택하고 그에 대응되는 펩티드 질량 값의 집합을 저장한 파일을 입력하면 펩티드 질량검색도구에 대한 사용자 인터페이스가 제시된다. 이는 검색기에 대한 질의어를 생성하는 단계이다. 이 인터페이스에서 요구하는 파라미터를 결정하면, 다수의 질량검색도구로 질의를 보내기 위하여 각 검색도구에 맞는 질의어를 생성한다.

질량검색도구들이 요구하는 파라미터는 서로 일치하는 것, 상이한 것, 경우에 따라서는 같은 의미의 파라미터 값의 단위가 틀린 것이 있다. 이에 대한 정보를 관리하여 각 인식도구에 적합한 질의어를 생성하고, 질의어에 대한 응답을 검색하기 위하여 각 도구들에 대한 정보가 필요하다. 그 정보의 종류로는, 검색할 데이터베이스, pI 범위, Mw 범위, 펩티드질량, 오차의 범위, 오차에 대한 단위, 최소일치 개수, 검색할 종 등 약 30개가 있다. 이를 위하여 메타데이터관리자는 단백질 인식도구들이 요구하는 파라미터에 대한 정보를 관리한다. 이러한 과정은 디지털라이브러리의 연합에서 각 디지털라이브러리에 대한 질의어로 번역하는 과정과 유사하다.

다음 단계는 질의결과를 통합하는 융합(fusion)단계이다. 이 단계에서는 각 질량검색도구들에게 동시에 보내진 다수의 질의어에 대한 응답을 취합하여 가장 가능성이 높은 단백질을 선택해야 한다. 각 도구들의 랭킹 알고리즘은 서로 상이하여, 주어진 질량 값에 대응되는 단백질의 순위가 서로 상이한 경우가 많이 나타난다. 검색결과에 대한 순위를 통합하기 위하여 각 검색도구들이 제공하는 정보를 관리해야 한다. 이 정보에는 순위, 점수, 일치하는 펩티드 개수, 종, 단백질 이름, 검색된 단백질에 대한 정보를 저장하고 있는 데이터베이스의 접근번호와 링크, pI, Mw, 각 질량에 대응되는 펩티드 서열 등이 있다.

이 중에서 일치되는 종을 우선적으로 조사하고, 다음으로 각 검색기에서 제시된 점수와 이론적으로 계산된 질량 값과 일치되는 비율을 조사한다. 만약 모든 질량 검색기의 결과가 일치하면서 일치되는 질량의 비율이 임계값을 넘으면 자동으로 해당되는 점에 대한 단백질이 결정되어서 2-DE 데이터베이스에 그 정보가 자동으로 저장된다. 그렇지 않으면, 각 질량검색기가 제시한 상위 세 개의 결과를 묶어서 부수적인 정보와 함께 사용자에게 제시하여 판단하게 한다. 이때 사용자는 검색도구에 대한 질의를 수정하여 다시 질의를 보낼 수 있다.

한 점에 대한 단백질이 인식된 이후에는 그 단백질에 관련된 정보를 얻기 위하여 해당 단백질에 대한 정보를 관리하는 단백질서열 데이터베이스, 대응되는 DNA 데이터베이스, 문헌정보데이터베이스들에 동시에 접근하여 필요한 정보를 모아서 2-DE 데이터베이스에 자동으로 저장된다. 각 데이터베이스가 제공하는 정보는 미리 약속된 코드로써 구조화되어 있기 때문에 필요한 정보를 추출하기가 비교적 쉽다. 이 과정에서는 필요한 정보만 추출하여 저장하기 때문에 검색결과에 대한 순위결정을 할 필요는 없다. 이 정보는 자체 구축한 임상병리데이터베이스와 연동해서 질병특이 단백질 발굴에 활용된다.

5. 구현

질량검색도구로는 MS-Fit(http://prospector.ucsf.edu/ucsfhtml3.4/msfit.htm), PeptIdent(http://www.expasy.ch/tools/peptident.html), PeptideSearch(http://www.mann.embl-heidelberg.de/Services/PeptideSearch/FR_PeptideSearchFormG4.html)를 사용하였다. 이들 도구들이 사용하는 질량데이터베이스는 주로 SWISS-PROT, TrEMBL (http://www.expasy.ch/sprot/)이다.

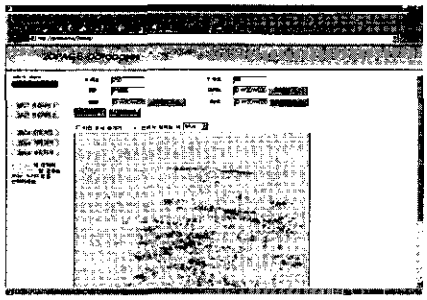

그림 4. 자동화 도구의 초기입력자료와 마스터 gel 이미지

그림 4는 개발된 도구의 초기화면이다. 사용자가 최초에 마스터 2-DE gel을 입력한 경우에는 인터페이스 하단에 그 이미지가 나타난다. 사용자가 이미지 상의 특정 점을 클릭하고 그 점에 대응되는 펩티드질량 값을 저장한 파일을 올리면 단백질이 명백히 인식된 경우에는 자동으로 데이터베이스에 그 내용이 구축된다. 그렇지 않은 경우에는 그림 5와 같은 인터페이스가 제시된다.

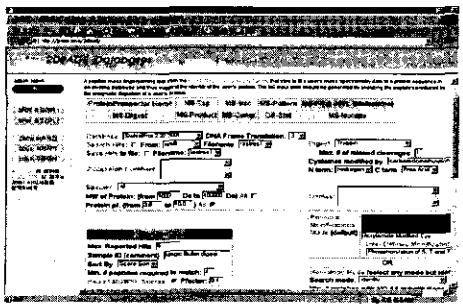

그림 5. MS-Fit를 이용한 매개변수 입력 인터페이스

그림 5의 인터페이스에서 제공할 매개변수는 초기에는 미리 고정되기 때문에 초기에 주어질 필요는 없다. 이는 단백질 인식결과가 명확하지 않을 때 사용자가 입력매개변수를 조정하여 질의어를 변경시킬 때 사용된다.

실험에서 자동으로 단백질 인식하기 위하여, 상위 두 개의 단백질에 대한 일치하는 질량 값의 비율의 차이에 대한 임계값을 10%이상으로 했을 때 안정적인 단백질 인식이 되었다. 그리고 약 10% 정도는 사용자의 판단이 필요하였다. 그림 5에서 제시된 인터페이스는 MS-Fit가 제공하는 것인데, 다른 인식기에서 필요로 하는 매개변수를 모두 포함하고 있기 때문에 그대로 사용할 수 있다. 펩티드 질량에 대응되는 단백질이 인식되면, SWISS-PROT과 TrEMBL의 단백질 서열 데이터베이스에 접근하여 그 단백질에 대한 정보를 수집하였고, GenBank에 접근해서 그 단백질에 대응되는 DNA 정보를 연결하였다. 그리고 해당 단백질에 관한 문헌정보를 수집하기 위하여 MedLine(http://www.ncbi.nlm.nih.gov/PubMed/)을 사용하였다. 본 도구의 구현에 사용된 언어는 ASP와 Java이다.

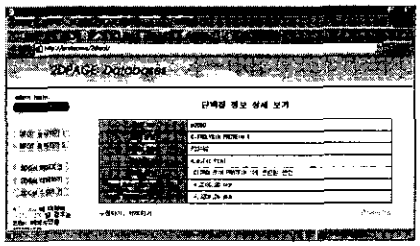

그림 6. 자동 구축된 2-DPAGE 데이터베이스 검색결과

그림 6은 2-DE 이미지 상의 한 점을 클릭했을 때, 구축된 2-DE 데이터베이스로부터 검색된 결과이다. 그 내용은 SSP번호, 단백질명, pI, Mw, SWISS-PROT 엑트리, GenBank 엔트리, 질량 스펙트럼 데이터, 관련문헌 등인데, 생물학적 요구에 따라서 그 정보를 쉽게 추가할 수 있다.

6. 결론 및 향후과제

수백 개의 프로테옴 라이브러리를 통합하는 문제는 프로테옴 연구에서 중요한 문제로 대두되었다. 통합의 유형은 다양하게 설계될 수 있는데, 본 논문에서는 다수의 프로테옴 관련 데이터베이스를 통합함으로서, 생물학적 실험에 의해서 생성된 펩티드 질량 값의 집합으로부터 2-DE 데이터베이스를 자동으로 구축하고 사용자에게 2-DE gel 이미지 상에 펼쳐진 단백질에 대한 정보를 자동으로 검색할 수 있도록 하는 도구에 대하여 논하였다.

프로테옴 라이브러리의 통합문제는 정보검색분야에서 활발히 연구되어 온 디지털라이브러리의 연합분야와 기술적인 면에서 공통적인 부분이 많이 있지만, 사용자가 원하는 서비스의 유형이 서로 다른 면이 많이 있기 때문에, 앞으로 다양한 유형의 통합서비스에 대한 많은 연구가 필요하다.

참고문헌

[1]. C. Hoogland, et al. "Make2ddb: a simple package to set up a 2-DE databaseon the WWW," Electrophoresis, 18, 1997

[2]. J. E. Celis, et al, "Human and Mouse Proteomic Databases: Novel Resources in the Protein Universe," FEBS Letters, 430, 64-72, 1998

[3]. Luis Gravano, et al. "Starts: Stanford proposal for internet meta-searching.", In Proceedings of the 1997 ACM SIGMOD, 207--218, May 1997

[4]. M. R. Wilkins(Editor), et al, Proteome Research : New Frontiers in Functional Genomics (Principles and Practice), Springer Verlag, 1997

[5]. http://www.expasy.ch/sprot/

[6]. P. F. Lemkin, "2DWG meta-databases of 2D electrophoretic gel images on the internet", Electrophoresis, 18, 2759-2773, 1997

[7]. P. F. Lemkin et al, "The Protein Disease Database of human body fluids," Applied and Theoretical Electrophoresis, 5, 55-72, 1995