DBMS 기반 단백질 식별 시스템의 설계 및 구현

이진관[0]*, 오석준**, 최은선*, 류근호*
*충북대학교 데이터베이스연구실, **바이오 정보기술 연구센터
*(jklee, eschoi, khryu)@dblab.chungbuk.ac.kr, **sjaugh@bi.snu.ac.kr

Design and Implementation of DBMS-based Protein Identification System

Jin Kwan Lee[0]*, Sirk June Augh**, Eun Sun Choi*, Kuen Ho Ryu*
*Database Lab. , Chungbuk National Univ., **Center for Bioinformation Technology

요 약

Human Genome Project의 완성을 전후로 이루어진 생물학적 분석도구의 발달과 서열 데이터베이스의 축적으로 단백질 분석은 괄목할 만한 성장을 하였다. 단백질 분석에 사용되는 가장 중요한 단백질 식별(Identification)을 위한 도구들은 많이 개발되어 왔으나 기존의 도구들은 파일 기반으로 폭발적으로 증가하는 단백질 데이터를 효율적으로 관리하고 실험자들에게 빠르고 정확한 검색결과를 제공하는데 한계를 보여주고 있다. 우리는 SWISS-PROT flatfile을 분석하여 관계형 데이터베이스로 구축하고, 단백질 식별에서 가장 많이 사용하고 있는 '질량분석 후 데이터베이스 검색' 방법을 사용하는 시스템을 DBMS 기반으로 설계하였으며, 다양한 실험조건과 절차에 의해 얻은 단백질 조각의 질량 값과 이론적인 계산에 의해 얻은 값을 비교하여 실험에 쓰인 단백질 조각과 일치하는 것을 단백질 데이터베이스로부터 검색할 수 있는 도구를 설계하고 구현하였다.

1. 서론

인간 유전자 지도의 완성을 계기로 인간 유전자에 대한 기능을 본격적으로 연구할 수 있는 이른바 포스트게놈 시대가 열렸다. 현재 18종 생물체의 게놈 구조가 완전히 해독되었고 지금도 생물체의 게놈 구조에 대한 해독 작업은 계속되고 있다[1].

그러나 아무리 염기서열을 밝혀낸다 하더라도 이들의 기능을 알지 못하면 큰 의미가 없다. 이렇게 유전자의 기능을 밝히는 총체적인 연구분야를 "Functional Genomics"(유전체 기능 분석학)라고 정의 하며, 여기에는 DNA나 RNA를 시료로 하여 수행하는 Genomics 와 단백질을 대상으로 연구하여 유전자 기능을 밝히는 Proteomics 그리고 이 두 분야를 지원하는 생물정보학 또는 계산 생물학으로 각각 구분된다.

DNA 와 RNA 를 사용하여 유전자의 기능을 밝히는 Genomics 가 현재까지는 유전체 기능 분석에서 주로 사용되었지만, 단백질 염기서열의 축적 및 단백질 분석도구의 발전은 점차 Proteomics 의 중요성을 인식하게 하였다. 이것은 Genomics 가 유전체 염기 서열만을 가지고 유전체의 기능을 밝히려 하는데 실제 유전체의 기능은 발현된 단백질의 단백질합성후 변형(post-translational modification)에 달려 있어서, 최종적으로 완벽한 모양이 갖추어진 단백질을 분석하지 않고는 그 유전자의 세포내 기능을 알 방법이 없기 때문이다. 우리는 단백질의 신규성과 기능을 연구하는데 필수적인 도구인 단백질 식별 시스템을 데이터베이스 관리 시스템을 사용하여 구현하였다.

이 논문에서는 Proteomics연구의 시작점이라 할 수 있는 단백질 식별 시스템의 설계 및 구현을 다루는데 2장에서는 관련 연구로 2DE(2D gel Electrophoresis)와 MS(Mass Spectrometry) 그리고 단백질 데이터베이스에 대해 알아보고, 3장에서는 식별 시스템의 설계 및 구현을 설명하고 4장에서는 결론 및 향후연구를 논하였다.

2. 관련연구

2.1 2DE(2D gel Electrophoresis)

2D 전기영동 기술이라 불리는 이 기술은 단백질의 pH 등전점(pI)으로 나타내는 net charge에 따라 1차 분석을 한 후(1D), 이어서 분자량에 따라 분리하는 (2D) 방법이다[2]. 그림 1은 sample에 대한 2DE의 결과를 보여주는 그림이다.

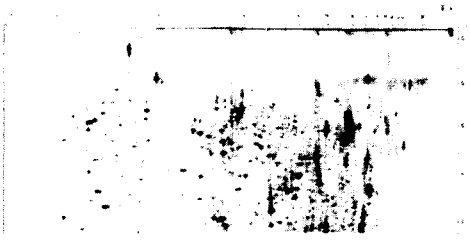

그림 1 2D gel electrophoresis

이 기술을 사용하면 분석 대상 단백질 샘플을 세포의 특정 생리조건에 따라 재조하여 특징 spot에 대한 분리 양상을 지도로 구성할 수 있다. 여기서 나타난 연구대상의 특징 spot의 단백질을 적당한 크기로 자르고, 이것을 trypsin을 사용하여 단편조각들로 만든 후 아래에서 설명할 MALDI-TOF등의 단백질 질량분석

기로 분석하여 아미노산 서열을 결정하고, 이를 바탕으로 단백질이나 genome database를 bioinformatics tool로 찾아 이 단백질의 정체를 확인하는 연속된 과정이다.

2.2 MS(Mass spectrometry)

단백질의 화학적인 분자량 측정으로 MALDI (Matrix-assisted Laser Desorption/Ionization) [3]와 ESI (Electrospray Ionization)[4]가 있다. MALDI는 2D-gel 상에 분리된 단백질을 trypsin으로 가수 분해하여 얻은 단일 전하를 띤 이온화된 peptide를 분석하는 기법이다. ESI 방법은 정제된 시료를 이온화 시켜 분석하는 방법으로 column chromatography와 연결시켜 사용할 수 있다. 효과적인 MALDI 분석을 위해서는 질소레이저를 사용, 분석대상물질의 비행시간을 측정하여 각각의 질량을 분석하는 TOF형이 사용된다.

2.3 SWISS-PROT

대표적인 단백질 데이터베이스로 SWISS-PORT, NCBI-nr, PDB, PIR등이 있으며 주로 단백질 데이터에 대한 annotation 정보를 제공하는 것과 단백질 구조를 제공하는 것으로 분류된다. 단백질 데이터에 대한 annotation 정보를 제공하는 데이터베이스 중 가장 많이 이용되는 것이 SWISS-PROT이다.

SWISS-PROT[5]은 다음과 같은 annotation 을 제공한다.

- 단백질의 기능에 대한 기술
- 도메인 구조
- post-translational modifications
- Variants

최소한의 중복과 cross-references를 통해 다른 단백질 서열 데이터베이스와의 높은 수준의 통합을 제공하며 sprot39 배포이후로는 filatfile이 배포되지 않았다.

그림 2는 플랫파일의 sample 이다.

```
ID   100K_RAT      STANDARD;      PRT;   889 AA.
AC   Q62671;
DT   01-NOV-1997 (Rel. 35, Created)
DT   01-NOV-1997 (Rel. 35, Last sequence update)
DT   15-JUL-1999 (Rel. 38, Last annotation update)
DE   100 KDA PROTEIN (EC 6.3.2.-).
OS   Rattus norvegicus (Rat).
OC   Eukaryota; Metazoa; Chordata; Craniata;
     Vertebrata;Euteleostomi;
RN   [1]
RP   SEQUENCE FROM N.A.RC   STRAIN=WISTAR;
     TISSUE=TESTIS;
RX   MEDLINE; 92253337.RA   Mueller D., Rehbein M., Baumeister
     H., Richter D.;
RT   small nuclear ribonucleoprotein particle (snRNP).";
RL   Nucleic Acids Res. 20:1471-1475(1992).
CC   -!- FUNCTION: E3 UBIQUITIN-PROTEIN LIGASE WHICH
     ACCEPTS UBIQUITIN FROM
CC   AN E2 UBIQUITIN-CONJUGATING ENZYME IN THE FORM
     OF A THIOESTER AND
DR   EMBL; X64411; CAA45756.1; -.
KW   Ubiquitin conjugation; Ligase.
FT   DOMAIN      77     88      ASP/GLU-RICH (ACIDIC).
SQ   SEQUENCE 889 AA; 100368 MW; ABD7E3CD53961B78
     CRC64;
     ADAVFSAMDL AFAVDLCKEE GGGQVELIPN GVNIPVTPQN
     VYEYVRKYAE HRMLVVAEQP LHAMRKGLLD VLPKNSLED
```

그림 2 SWISS-PROT 플랫파일

플랫파일의 데이터필드의 의미를 간단히 설명하면 다음과 같다.

ID: 엔트리 이름과 서열 길이
AC: 접근 번호이다. 주요 접근 번호와 보조 접근 번호
DT: 데이터가 생성된 날짜, 마지막으로 갱신된 날짜, annotation이 마지막으로 변경된 날짜
DE: 엔트리에 부여된 gene 이름
OS: 종(Species)
OC: 엔트리가 속하는 계통
RN: 참조번호
RP: 참조 위치
RC: 참조에 대한 코멘트
RX: 참조에 대한 참조
RA: 참조의 저자
RT: 참조의 제목
RL: 참조의 위치
CC: 엔트리에 대한 코멘트
DR: 다른 데이터베이스에 대한 참조
KW: 키워드
FT: 서열 각 부분에 대한 설명
SQ: 서열의 길이, 분자량, 서열

위의 각 항목은 필수항목(ID, AC, DT, DE, OS, OC, SQ)와 엔트리에 따라서는 표시되지 않는 항목(필수 항목 이외의 항목)으로 구분된다.

3. 단백질 식별 시스템의 설계 및 구현

3.1 시스템 구조

질량분석을 통해 얻은 단백질 질량을 가지고 단백질 데이터베이스를 검색하여 단백질을 식별하는 시스템의 절차는 그림 4와 같다.

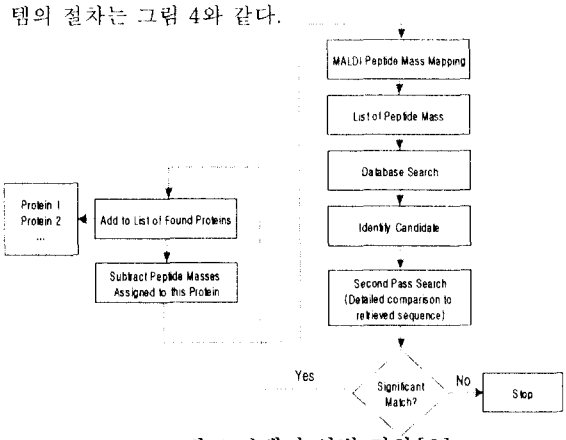

그림 4 단백질 식별 절차[6]

위와 같은 단계를 따르는 시스템의 세부 구성요소들을 보면 크게 데이터베이스 구축과 실제 단백질 식별과정에서 사용자가 입력한 데이터를 가지고 데이터베이스를 검색하는 검색기 그리고 검색된 결과를 평가하여 가장 신뢰도가 높은 데이터들 만을 보여주는 식별 결과 평가 모듈로 나눌 수 있다. 각각을 구성하는 요소들은 그림 5와 같다.

3.2 SWISS-PROT 플랫파일을 사용한 데이터베이스 구축

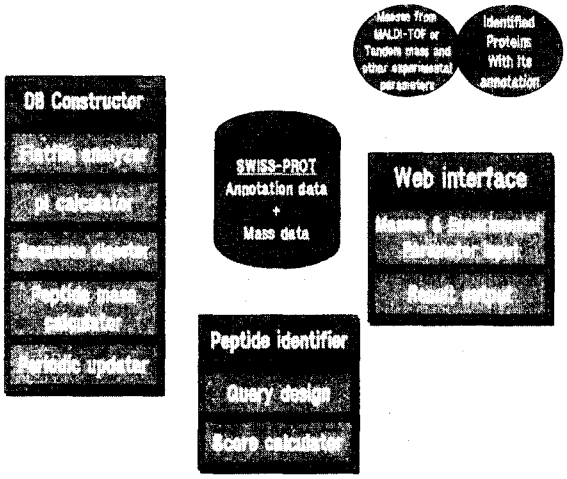

그림 5 시스템 구조

관련연구에서 보았듯이 SWISS-PROT 플랫파일은 하나의 단백질 엔트리에 대해서 다음과 같은 annotation 정보들을 갖는다; ID, DT, DE, OC, RN, RP, RX, RT, RL, CC, DR, KW, FT, SQ. 이 정보와 서열로부터 계산한 이론적인 질량값을 갖는 릴레이션을 설계하고 relationship을 표현한 것이 그림 6이다.

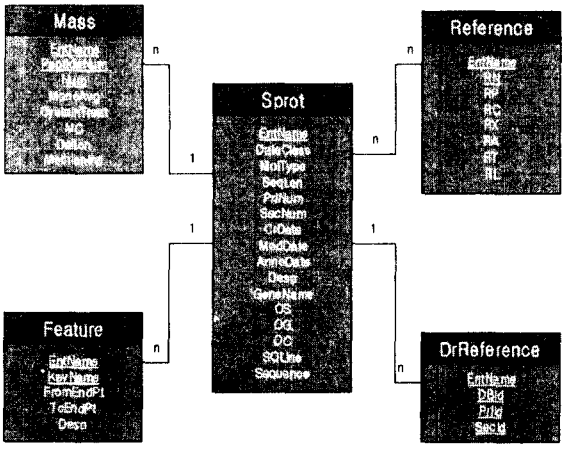

그림 6 Relationships for SWISS-PROT 플랫파일

위 그림에서 릴레이션 Sprot, Reference, DrReference, Feature는 플랫파일을 분해한 것이고 Mass는 아미노산 서열로부터 Digestion Enzyme으로 Trypsin을 사용하였을 때 얻어지는 가수분해된 peptide들에 대하여 실험에 사용될 각각의 경우에 대하여 얻어지는 질량값에 대한 것이다.

3.3 시스템의 수행

우리는 이 시스템을 다음과 같은 환경에서 구현하였다.

- DBMS: Oracle 8i
- 운영체제: SunOS 5.7
- 프로그래밍 환경: JDK 1.2

JDBC를 통해 데이터베이스와 연결하고 JSP를 사용해 웹을 통해 단백질 식별 엔진과 사용자가 연결된다. 실험 데이터를 통해 수행한 결과 ExPASy에서 제공하는 PeptIdent와 유사한 결과가 나왔으며 상이한 결과는 ExPASy에 새로 갱신된 데이터이다.

4. 결론 및 향후연구

지금까지 단백질 식별을 위한 소프트웨어들[7]은 10여가지가 개발되어 왔으나 파일 기반이기 때문에 폭발적으로 증가하는 생물데이타를 처리하는데 그 한계를 보여주고 있다. 점차로 생물데이터 자체를 사유화하고 공개하지 않는 추세인 것을 감안한다면 국내에서 만들어지는 단백질 데이터를 자체적으로 기존의 데이터베이스에 추가하여 구축하는 것이 바람직하다고 본다. 이러한 취지로 우리는 단백질 데이터베이스 중의 하나인 SWISS-PROT을 관계형 데이터베이스로 구축하였고 이 데이터베이스를 검색하여 단백질의 신규성과 기능을 규명하기위한 도구인 단백질 식별 시스템을 설계하고 구현하였다.

단백질 식별 시스템에서 중요하게 부각되고 있는 것은 검색의 결과에 대한 평가 시스템이다. 검색 결과에 대한 평가를 위해 Scoring 기법[8]을 사용하는데 Scoring 기법에는 매치되는 질량의 개수를 세는 방법, 평균에 의한 방법, Bayesian 방법이 있다. 현재에는 Bayesian 방법에 기초한 MOWSE 알고리즘이 가장 정확한 평가 방법이라 한다. 현재 우리가 구현한 시스템에서는 매치되는 질량을 세는 방법을 사용하는데 MOWSE 알고리즘을 사용하여 평가 기능을 향상할 것이며 NCBI-nr등과 같은 다른 데이터베이스도 추가로 구축할 것이다..

참고문헌

[1] 백용기, 프로테오믹스의 연구방향과 국내 프로테오믹스 연구의 활성화, BioWave, 2000, 7. Vol 2. Num 7.

[2] Dr James R. Jefferies, Parasitology Group, etc. 2D GEL ELECTROPHORESIS TUTORIAL, http://www.aber.ac.uk/~mpgwww/Proteome/Tut_2D.html

[3] Karas, M., Bachmann, D., Bahr, U., Hillenkamp, F., International Journal of Mass Spectrometry and Ion Processes; 78, 1987, pp53-68

[4] Fenn, J.B., Journal of the American Society for Mass Spectrometry, 4, 1993, pp524-535

[5] Amos Bairoch and Rolf Apweiler:The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000, Nucleic Acids Research, 2000, Vol.28, No1

[6] Ole N. Jensen, Alexander V. odtelejnikov, and Matthias Mann: Identification of the Components of Simple Protein Mixtures by High-Accuracy Peptide Mass Mapping and Database Searching, Analytical Chemistry, Vol. 69, No. 23, December 1, 1997.[8] Ronald C. Beavis, David Fenyo, Database searching with mass spectrometric information, Proteomics: A Trend Guide, July 2000