분산 환경에서 Peptide Mass Mapping에 의한 단백질 검증 시스템 설계 및 구현

신민수O*, 허철구**, 임소형***,김도완*
배재대학교 정보통신공학과*, 생명공학연구소 유전체연구센터**,이화여대 분자생명공학부***

The Protein Identification system Design and Implementation by Peptide mass mapping in Distributed Environment

Min-Soo Shin*, Chul-Gu Hur**, So-Hyung Lim***, Do-Wan Kim*
Dept. of Information and Communications Engineering, PaiChai University*
Genome Research Center, Korea Research Institute of Bioscience & Biotechnology**
Dept. of Molecular and life Science, Ewha somans University***

요 약

오늘날 단백질 정보 분석은 HGP(Human Genome Project)이후 Post-genome 시대를 맞이하면서 매우 중요한 분야로 인식되고 있다. 이 단백질 정보를 이용하는 응용은 Discovery of Protein Structure/Function Relationships, Evolutionary Relationships, 3D Modeling 등 많은 분야에서 활용되어진다. 여러 가지 분야들 중에서 특히 단백질 구조 분석을 위한 많은 다양한 소프트웨어들이 출현되고 있다. 하지만 복잡하게 얽혀 있는 단백질들을 검증하기 위해서 Mass Spectrometry에서 발생되는 Peptide Masses의 정보들을 이용할 수 있다. 이에 본 논문에서는 Mass Spectrometry에서 생성된 Peptide Mass Map을 이용하여 기존의 단백질 Database에 있는 단백질들과 비교하는 자동화 단백질 검증 시스템 설계 및 구현에 관한 연구내용을 담고 있다. 이 시스템은 3-계층 중심으로 개발이 이루어지며 이 기종 시스템과의 원활한 통신, 다중 계층의 환경에 있는 각 객체들간의 통신을 위해서 RMI 기반의 미들웨어를 활용하기로 한다.

1. 서 론

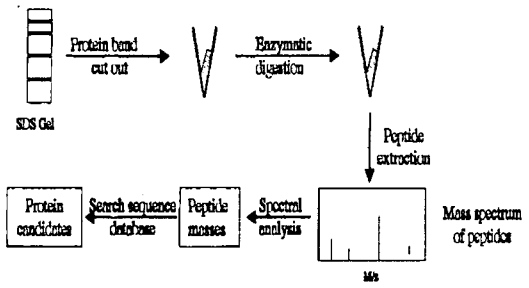
과학의 발달사를 되돌아볼 때 한 분야의 기술 혁신이 학제간의 상호적인 작용에 의하여 촉진되어 온 많은 예를 볼 수 있다. 광학이 생물학의 발전에, 화학이 원자물리에, 천문학이 입자물리에, 그리고 전자공학이 컴퓨터과학의 발전에 획기적인 기술을 제공하였으며, 그 역 기여 또한 많았다. 생명공학의 기술혁신은 컴퓨터의 간접적인 효과를 받아왔다. 한 예로 지난 10년에 시작한 게놈 사업은 15년에 걸쳐 끝난다고 했지만 컴퓨터 기술 분야의 절대적인 역할에 의해서 10년으로 단축되는 성과를 이루어냈다.[1] 많은 연구에 의해 수많은 생물학적 서열 데이터가 수집되어 데이터베이스로 구축되었다. 최근에 축적된 데이터베이스를 관리하고 분석함으로써 유용한 지식을 얻어내어 생명 과학의 여러 문제에 활용하고 있다. 즉 축적된 데이터의 분석을 통해 DNA 서열 분류, DNA 서열에서의 단백질 코딩 영역 분류, 분자의 구조와 기능의 예측, 진화 역사의 구축 등의 생명 현상 탐구가 가능하게 되었다. 본 논문은 특정 단백질을 화학적인 반응 또는 특정 protease에

의해서 분해해서, MALDI(Matrix-assisted laser desorption/ionization) 또는 ESI(Electrospray ionization)와 같은 실험방법에 의해서 산출되는 peptide masses들과 데이터베이스내의 각 단백질을 특정한 Enzyme Agent를 이용하여 분해하여 산출된 Peptide Masses과 비교하여 여러 가지로 복잡하게 구성된 단백질들 속에서 특정 단백질을 검증할 수 있는 자동화된 단백질 검증 시스템을 제시 하고자 한다. 2장은 실험적인 방법론에 대해서 기술하며, 3장은 실제 단백질 검증 시스템의 설계 및 구현 부분을 기술한다. 마지막으로 4장의 결론 및 향후 발전 방향에 대해서 논의할 것이다.

2. 관련 연구

현재 단백질을 검증하는 방법은 여러 종류들이 있다. Peptide의 서열 tag 정보를 이용하여 하는 방법, tandem mass spectra을 정보를 이용하는 방법, 그리고 가장 일반적인 방법으로 본 논문에서 활용한 peptide mass mapping을 이용한 방법이 있다.[2]

그림1은 개괄적인 흐름도를 이용하여 단백질 검증과정을 설명하고 있다. SDS Gel이 복잡하게된 혼합된 protein들을 분해하고, 이렇게 분해된 단백질들의 결과들은 Matrix-assisted laser desorption/ionization mass spectrometry(MALDI-MS) 또는 Electrospray ionization mass spectrometry(ESI-MS)에 의해서 Peptide의 Mass Spectrometry를 형성하게 된다. 검



[그림 1 단백질 검증을 위한 절차를 나타내는 흐름도]

색은 Peptide 레벨에서 이루어지며, 적절한 알고리즘을 바탕으로 실험적으로 발생된 Peptide Masses는 단백질 데이터베이스 안에 있는 각 단백질을 특정 Enzyme Agent를 이용하여 Peptide들로 분해한 다음 각 Peptides들의 Mass값을 계산한 다음 이미 실험적으로 발생된 Peptide Masses들과 비교한다. 이러한 비교에 의한 평가를 기반으로 단백질을 검증하게 된다. 단백질 검증을 위한 Peptides Mapping 방법은 대체적으로 다른 방법들에 비해서 빠른편인데, Mass Spectra와 같은 정보들을 빨리 알 수가 있기 때문이다. 예를 든다면 human hemoglobin α는 Trypsin이란 Enzyme agent로 분해되어 14(728.8, 460.5, 531.6, 1529.6, 1071.3, 1834.0, 397, 146.2, 2997.3, 287.4, 817.94, 2967.5, 1252.5, 337.4)개의 Peptides Masses들을 예상할 수 있다. 실험적으로 Peptide Masses들은 Mass Spectrum에 의해서 획득할 수가 있다. 여기서는 모두 7(728.8, 1529.6, 1834.0, 287.4, 1252.5, 650.3, 1345.0)개의 Peptides Masses들로 구성이 되어 있다. 즉 마지막 두 개는 hemoglobin 서열로부터 생성된 Peptide들과 연관이 없다. 위의 7개의 관측된 Peptides Masses를 이용해서 단백질 데이터베이스 안에 있는 모든 단백질의 Peptide Masses들과 비교한다. 우리는 관측된 Peptide Masses와 단백질로부터 생성된 Peptide Masses들의 정합(matching)된 수를 기록한다. 위와 같은 방법을 이용한 단백질 검증은 데이터베이스 내의 하나의 단백질 서열과 실험적으로 생성된 데이터 사이에 최상의 정합(matching)을 결정하기 위한 스키마를 구성하게 되며, 이 시스템에서는 Peptide의 정합된 순으로 순위를 결정하였다.[3][4]. 본 단백질 검증 시스템은 위와 같은 알고리즘을 근거로 시스템을 구성하였다.

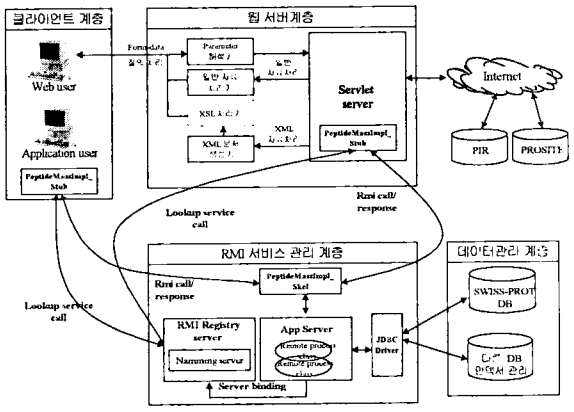
3. 시스템의 설계 및 구현

3.1 단백질 검증 시스템 구조 및 설계

본 시스템은 Peptide Mass Mapping을 기반으로 특정 단백질 검증하기 위한 시스템으로 일반 웹 사용자와 응용 기반의 사용자에 모두에게 서비스를 제공하도록 설계 및 구현이 되었다. 즉 기존의 2-Tier 모델이 가지는 클라이언트 유지보수의 문제점, 무거운 클라이언트 구조, 데이터베이스 서버의 과부하 초래, 객체의 재사용성의 어려움 등의 문제점을 다단계 분산 시스템으로 구성함으로써 서버가 가지는 과다한 부하를 조절해 주며, 클라이언트의 계속적인 변경에 대해서 서버와는 독립적인 형태를 취할 수가 있다.[5,6] 현재 이 시스템은 웹 사용자를 위해서는 Servlet으로 구현을 했으며, 애플리케이션 사용자를 위해서는 자바 애플리케이션으로 작성이 되어있다.

단백질 검증 시스템은 다단계 분산 시스템으로 구성이 되어 있고 현재 4단계의 계층 구조로 구분이 되어진다.

첫 번째 클라이언트 계층이다. 클라이언트 계층은 일반 웹 사용자나 응용 프로그래머가 속해져 있는 계층이다. 이 계층에서 웹 사용자는 Servlet server를 통해서 현재 자신이 원하는 서비스를 하는 서버가 어디에 있는지를 Lookup 서비스를 이용해서 찾



[그림 2 분산 환경에서 Peptide Mass Mapping에 의한 단백질 검증 시스템 구조도]

게 된다. 애플리케이션 사용자는 RMI registry 서버와 직접 통신을 함으로써 서비스를 받게 된다.

두 번째는 웹 서버 계층이다. 웹 사용자로부터 요청된 서비스가 존재하는 서버를 찾는 작업을 하며, 그 서버로부터 실제 서비스를 할 원격 객체의 위치를 전달받게 된다. 호출한 원격 객체에 실제 단백질 검증에 필요한 모든 인자값을 전달하게 되며, 연산 결과들이 다시 Servlet 서버에게로 전달된다. 이때 Servlet 서버는 웹 사용자에게 결과를 전달할 경우 두 가지 형태를 통해서 전달하게 되는데 하나는 일반 텍스터의 연산결과를 다른 하나는 어떤 하나의 단백질 정보를 제공하는 XML 양식의 결과를 전달한다. 이 XML 결과는 XSL 처리기에 의해 서버에서 HTML 형태로 전환하여 사용자에게 전달한다.

세 번째는 RMI 서버관리 계층이다. RMI 서버관리 계층은 여러 가지 일을 하고 있다.

▷ RMI registry 서버 :

각종 응용 서버들을 바인드를 했어 클라이언트 계층 또는 웹 서버 계층으로부터 요청되는 서버의 찾아 주는 역할을 한다.

▷ Amino_acid_residue 클래스 :

각 Amino acid의 Molecular weight의 정보들을 관리하는 클래스이며, Monoisotopic 값과 Average 값을 가지고 있으며 PeptideMassImpl 클래스에서 특정 Enzyme Agent에 의해서 산출된 Peptide의 Mass를 산출하기 위해 적용된다.

▷ PeptideMassImpl_Stub 클래스 :

Servlet 서버나 응용 프로그램에서 lookup 서비스를 이용하여 서버를 찾은 다음 실제 원격 객체의 메소드를 호출한다.

▷ Detail_info_class 클래스 :

PeptideMassImpl_Skel 클래스에서 마샬링을 이용해서 클라이언트나 웹 서버에게 전달되는 객체의 단위로 Peptide의 시작점, 끝점, 관련된 Peptide 서열들의 정보를 다루는 객체로써 Serialization으로 구현되어 있다.

▷ PeptideMass_class 인터페이스 :

RMI 시스템에서 가장 중요한 역할을 담당하는 인터페이스로써 클라이언트가 각종 원격 객체의 메쏘드를 호출할 수가 있다.

▷ Peptide_class 클래스 :

한 단백질에서 발생할 수 있는 모든 Peptides의 정보들을 총괄하는 클래스로써 역시 Serialization으로 구현되어 있다

▷ PeptideMassImpl 클래스 :

단백질 검증 시스템에서 데이터베이스에 있는 단백질들을 특정 Enzyme Agent에 의해서 분해하고, 이렇게 생성된 각각의 Peptide들의 mass 값을 계산한다. 또한 기존의 실험적인 Peptide Masses들과 정합(matching)하여 후보 단백질의 순위를 정해서 클라이언트 또는 Servlet 서버에 전달한다.

▷ PeptideMassImpl_Skel 클래스 :

응용 서버로부터 생성된 결과들을 서비스를 요청한 스텁에게 전달하는 역할을 한다.

▷ PeptideResult 클래스 :

SWISS-PROT DB나 인덱스 관리 DB로부터 정해진 결과들을 호출한다.

▷ ProteinIndexExtract 클래스 :

Servlet 서버에서 인터넷을 통해 다른 단백질 데이터베이스 시스템으로부터 생성된 자료들을 인덱스 관리 DB에 저장하기 위한 전달되어 온 자료들을 파싱한다.

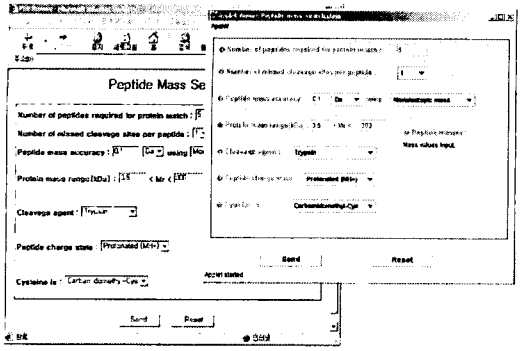
▷ DBConnectionImpl 클래스 : 응용 서버에 존재하는 각종 객체들이 SWISS-PROT DB나 인덱스 관리 DB에 접속해서 자료들을 추출할 수 있는 DB의 접속과 해제에 관련된 작업을 한다.

네 번째는 데이터관리 계층이다. 데이터관리 계층은 XML 기반의 단백질 정보들을 관리하는 SWISS-PROT DB와 다른 데이터베이스의 단백질에 관련된 link 정보들을 수집한 인덱스 관리 DB로 구성이 되어 있다.

3.2 단백질 검증 시스템 구현

3.2.1 실행화면

그림3은 실제 클라이언트 계층에서 단백질 검증 시스템을 위한 두 가지 사용자 인터페이스이다. 하나는 웹 기반으로 다른 하나는 자바 애플리케이션으로 작성했으며, 두 인터페이스 모두 동일한 작업 효과를 가진다.



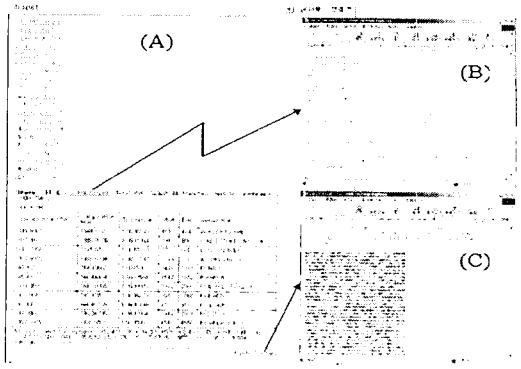
[그림 4 클라이언트 계층의 UI]

3.2.2 실행결과

그림4에서는 다음과 같은 정보들을 제공해 주고 있다. 우선 클라이언트 계층(A)에서 전달 받은 인자값(Matching peptide number, Mass accuracy, Mass range, Experimentally peptide masses, Missed cleavage site number, Cleavage agent etc)을 기반으로 원격 메소드를 통해서 얻어진 최종 결과물이다. 우선 실험적인 Peptide Masses와 데이터베이스의 단백질에서 파생된 Peptide Masses와 비교하여 정합된 수가 많은 순으로 순위를 결정하였다. 각 개별적으로는 링크 순위, 단백질 ID, 총 Molecular Weight , 실험적인 데이터와 일치하는 각각의 Peptide들에 관한 정보 등등이 있다. 단백질 ID의 링크(B)는 좀 더 세부적인 정보를 제공하며, 이것은 XML 기반의 Swiss prot DB에서 관련 정보들을 추출한다. 그 다음 Fasta format 의 링크는 단백질 서열간의 비교 또는 데이터베이스 내의 염기 서열간의 비교를 위한 일정한 양식을 제공한다.

4. 결론 및 연구 방향

본 연구에서는 Peptide Mass Mapping을 이용해서 단백질 검증하기 효율적인 방안으로 분산 환경에서 처리할 수 있는 시스템을 제안하였다. 기존의 2-Tier 모델이 가지는 클라이언트 유지보수의 문제점, 무거운 클라이언트 구조, 데이터베이스 서버의 과부하 초래, 객체의 재사용성의 어려움 등의 문제점을 3-Tier 계층의 분산 시스템으로 구성함으로써 서버가 가지는 과다한 부하를 조절하며, 각 계층간의 상호 연관성 및 독립성을 유지하도록 구현했다. 향후 연구 방향으로는 후보 단백질의 우선 순위 결정에 있어서 실험상의 Peptide masses와 계산된 Peptide masses들간의 정합 수를 기반으로 순위를



[그림 3 (A) 연산된 결과 리스트, (B) 특정 단백질에 관한 정보 제공, (C) Alignment 시스템에 사용될 Fasta format]

결정하는 것보다는 Scoring algorithm, Bayesian algorithm과 같은 고수준의 algorithm을 적용하여 좀더 신뢰성을 가질 수 있는 결과를 산출하는 좀더 지능화된 시스템으로 설계하고자 하며, 단백질 정보 분석을 더욱 자세히 하기 위하여 유전자 DB의 이종인 EST DB와 Mapping는 시스테으로 확장할 예정이다.

5. 참고문헌

[1] 김삼묘, "컴퓨터과학과 생명과학의 만남" 정보과학회지 2000년 8월

[2] Ronald C.Beavis, David Fenyö, "Database searching with mass spectrometric information"

[3] Wenzhu Zhang, Brian T.Chait, "ProFound: An Expert System for Protein Identification Using Mass Spectrometric Peptide Mapping Information" Anal. Chem. 2000, 72, 2482-2489.

[4] Matthias Mann, Peter Hojrup, "Use of Mass Spectrometric Molecular Weight Information to Identify Proteins in Sequence Database", Biological Mass Spectrometry, vol 22, 338-345(1993).

[5] George Reese, "Database Programming with JDBC and JAVA", O'Relly, 1996

[6] Copyright d-tec Distributed Technologies CmbH, 1998.

[7] Jan Eriksson, Brian T.Chait, and David Fenyö, "A Statistical Basis for Testing the Significance of Mass Spectrometric Protein Identification Results", Anal. Chem. 2000, 72, 999-1005

[8] Peter James, Manfredo Quadroni, Ernesto Carafoli, "Protein Identification by mass profile fingerprinting", Biochemical and Biophysical Research Communication, Vol 195, No 1, 1993.