기술혁신학회지 제2권 제3호
1999년 12월 pp. 94-106

분석기기지원을 위한 원격 데이터 분석 시스템 개발

Development of Remote Data Analysis System for the Joint Use of Equipments

권 경 훈* · 최 인 식** · 김 미 옥***

〈目 次〉

<Abstract>

In Korea Basic Science Institute (KBSI), the remote data analysis system is developed for the joint use of advanced equipments. This system enables the researchers to access the datas which are produced at KBSI and analyse them by Java program on the Web. Except Web browser such as Internet Explorer or Netscape Navigator, no additional softwares are required for analysing data. We have developed remote data analysis systems for five major equipments which KBSI supports for the researchers. The systems which are developed are those for NMR spectrometer, High Resolution Tandem Mass Spectrometer, Microscopic Imaging System, DNA Sequencer and Natural Radioactivity Measurement System. These programs work on any computer platform and any operating system, only if the internet is available. This remote data analysis system will be served as a part of Collaboratory, the remote collaborative system.

Key word : Remote Data Analysis, NMR, Mass Spectrometer, Microscopic Imaging System, DNA Sequencer, Natural Radioactivity Measurement System

* 기초과학지원연구소 과학전산실 선임연구원
** 기초과학지원연구소 과학전산실 책임연구원
*** 기초과학지원연구소 과학전산실 연구원

Ⅰ. 서 론

전세계적으로 인터넷의 활용이 보편화되면서 이전에는 방문, 또는 우편으로 처리하여 며칠씩 걸리던 작업들이 네트웍을 이용하여 수초내에 처리될 수 있게 되었다. 뿐만 아니라 컴퓨터의 성능이 급속히 발전하면서, 개개인의 PC에서도 대용량의 데이터를 빠른 속도로 처리할 수 있는 능력을 가지게 되었다. 네트웍과 고성능 컴퓨터는 다음 세기에나 가능하리라고 믿어졌던 서비스들을 실현하고 있다.

기초과학 분야도 이러한 발전에 있어서 예외는 아니다. 기초과학 분야에서 발 표되는 새로운 발견, 새로운 논문들은 학술지에 인쇄본으로 발표되기 수개월 전에 데이터베이스화되어 전세계 연구자들에게 배포된다. 공간적으로 멀리 떨어진 연구자들은 공동연구 시스템에 의해 응용프로그램, 데이터 등을 공유한다. Microsoft 사의 NetMeeting, NCSA (National Center for Supercomputing Applications) 의 Habanero 등의 소프트웨어는 원격지의 연구자와 화면의 윈도우, 응용프로그램을 공유함으로써 공동연구에 이용되고 있으며, 네트웍 속도의 증가, 네트웍 장비의 발전에 따라 보다 다양한 기능을 가진 소프트웨어들이 현재 개발 중에 있다. 한편 최근 제작되는 실험장비들은 인터페이스가 디지탈화되어 이전에는 수동으로 조작하던 작업들을 컴퓨터를 통해 더욱 세밀한 조작을 할 수 있게 하였으며, 컴퓨터를 통한 장비의 조작 메카니즘은 영역을 확대하여 네트웍을 거쳐 원격으로 장치를 제어할 수 있는 환경을 구축하게 하였다. 이러한 소프트웨어 및 하드웨어에서의 발전은 국가간, 연구소간의 공동연구 체제인 Collaboratory 라는 용어를 창출하였으며, 연구자간의 공간적 거리는 과학기술 연구에 있어서 더이상 제약조건이 되지 않는다. Collaboratory 시스템은 대형 고가 장비의 사용자 범위를 확대함으로써 장비의 활용도를 높이고 연구소간의 공동연구를 활성화하며, 공동연구에 필요한 여행비를 절감하는 효과를 가져온다.

미국 Oak Ridge National Laboratory (ORNL) 에서는 DOE 2000 project 참여자로서 1995년에 ORNL 의 수퍼 컴퓨터와 Sandia National Laboratory 의 수퍼컴퓨터를 ATM 으로 연결하여 하나의 코드를 분산 수행하는 시연을 보인 바 있으며, 또한 전자현미경의 원격제어 시스템을 개발하였다.(Parvin, 1997) 분석장비 분야의 원격 실험 시스템은 ORNL 외에도 University of California at San Diego 의 San Diego Microscopy and Imaging Resources (Mercurio, 1992), Environmental Molecular Science Laboratory (Bair, 1998), 일본의 금속재료 기술연구소 (http://inaba.nrim.go.jp/G7/remotelab2.html) 등의 여러 연구소에서 개발 운영 중에 있다. 한편 국가간의 협력이 필수적인 핵융합 연구개발 분야에서는 미국 General Atomics 사의 DIII-D (Casper, 1998), MIT 대학의 Plasma Science and Fusion Center의 Alcator C-Mod 의 핵융합 장치 (Fredian, 1999)를 원격으로 제어할 수 있다. 현재 개발된 원격 공동연구 시스템들을 살펴보면, 실험 장치의 제어에 있어서 신속한 처리

를 요구하는 부분, 실험에 큰 영향을 미치는 부분들은 장비를 보유한 연구소에서 직접 수행하며 그렇지 않은 기능들은 원격 처리를 허락하는 방식으로 기능을 이중화하고 있다. 그러나, 새로운 네트웍 기술 및 장비가 개발됨에 따라 네트웍을 통한 데이터 처리 속도는 급속히 개선될 것으로 예상되며, 가까운 장래에는 실험장비 제어를 위한 거의 모든 작동을 원격으로 수행할 수 있을 것이다. 이러한 세계적 동향에 발맞추어 기초과학지원연구소에서는 원격 공동연구시스템의 구축을 계획하고 있으며, 이에 대한 기초 단계로 원격 데이터 분석 시스템을 개발하였다.

기초과학지원연구소의 분석기기지원사업은 기초과학 연구에 필요한 고가의 첨단실험장비를 설치하고 대학 및 연구소에 산재한 연구자들이 이를 공동으로 활용하여 연구의 효율을 높임을 목적으로 한다. 이에 1989년 9월 이래로 국내 대학 및 연구기관들에 기기지원업무를 수행하여 왔다. 기기지원 신청이 접수되면 장비 담당자는 해당 시료를 분석하고, 그 결과물을 우편으로 발송한다. 결과물은 스펙트럼의 인쇄물, 분석 사진, DNA 서열 등으로서 실험장비에 연결된 컴퓨터의 소프트웨어에서 분석, 가공되어 분석신청자에게 보내진다. 하나의 실험에 대하여 스펙트럼을 확대하고 싶다든가 데이터에 어떤 변환을 실행하고자 할 때, 분석신청자는 장비담당자에게 이를 알려서 결과물을 다시 받아야 한다.

자기공명팀의 핵자기공명 (Nuclear Magnetic Resonance : NMR) 분광기 스펙트럼의 경우는 데이터 파일을 FTP 로 서비스하기도 한다. Anonymous FTP 서버에 저장한 데이터 파일을 연구자들은 각자의 연구실에서 인터넷을 통해 다운로드 받는다. 이렇게 받은 데이터 파일은 binary 파일이므로 이를 읽기 위해서는 소프트웨어가 필요하다. 연구자는 다운로드 받은 데이터 파일을 자신이 가지고 있는 분석 소프트웨어에서 분석한다.

이와 같이 종래의 분석기기지원 서비스는 분석신청자들이 원하는 형태의 결과물을 얻기 위해서 장비 담당자에게 전화로 또는 방문하여 이를 설명하고 요청하거나, 데이터 파일을 읽고 분석할 수 있는 소프트웨어를 개별적으로 구입해야만 했다. 실험 데이터를 분석하고 처리하는 소프트웨어는 대부분 고가이어서 연구자들이 개별적으로 구입하기에 어려우므로, 소프트웨어를 보유하지 못한 연구자들은 실험 데이터를 다운로드 받는다해도 이를 활용할 수 없으며 또다시 장비 담당자에게 원하는 형태의 결과물을 요청해야 한다.

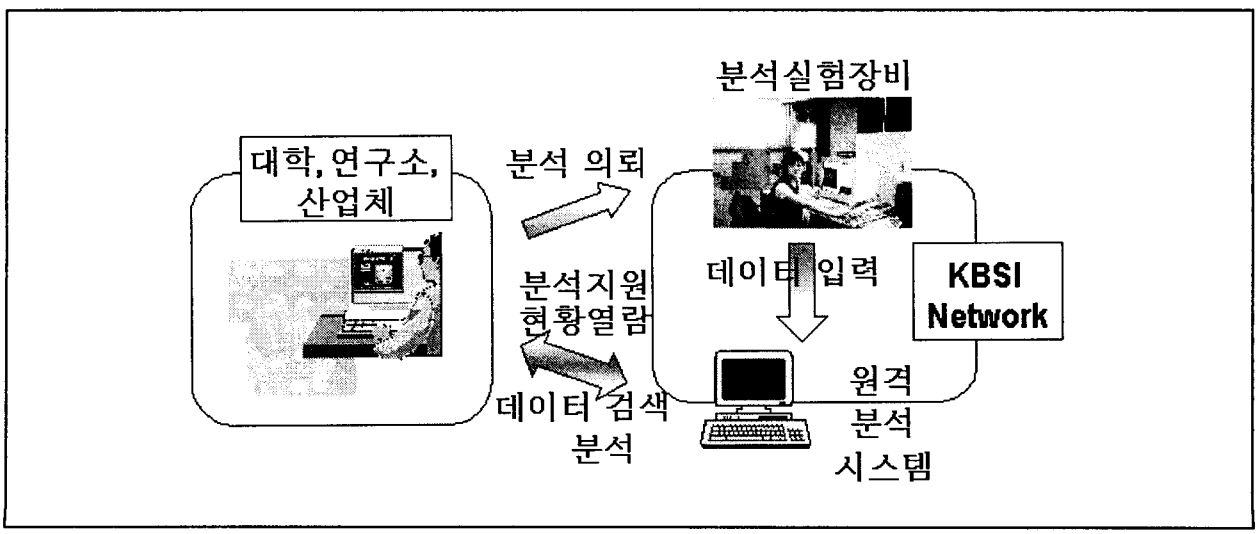
이처럼 분석 신청자들이 원하는 형태의 결과물을 받기위해 실험 결과물을 가공하는 데에 있어서 과정이 번거로운 단점을 개선하고자 기초과학지원연구소의 과학전산실에서는 원격 데이터 분석 시스템을 개발하였다.(URL 주소 http://polar.kbsi.re.kr/data_server/) 원격 데이터 분석 시스템은 분석신청자가 자신의 실험 데이터를 Internet Explorer 나 Netscape 와 같은 웹 브라우저를 이용하여 검색하고 분석한다. 1997년에 실험데이터를 GIF 형태의 그래프로 변환하여 웹브라우저에서 보여주는 방식의 데이터 서버를 구축(권경훈, 1997) 한 이후에, 자바 프로그램을 통해 데이터를 다양하게 분석,

가공할 수 있도록 업그레이드하였다. 이 시스템은 실험 데이터를 데이터베이스에 입력하고 분석 프로그램을 제공하는 방식으로 운영하는데, 이렇게 구축된 실험데이터 DB는 실험결과물의 체계적인 관리 및 저장에도 기여하는 바가 크다.

Ⅱ. 시스템 개요

[그림 1]은 원격 데이터 분석 시스템의 구성도이다. 분석지원 실험장비로부터 원격 데이터 분석 시스템 서버에 입력된 데이터는 인터넷을 통해 대학, 연구소 및 산업체에 제공된다. 원격 데이터 분석 시스템은 데이터베이스 처리 시스템과 데이터 분석 시스템으로 구성된다. 데이터베이스 처리 시스템은 데이터를 DB에 입력하고 검색하는 시스템으로 기초과학지원연구소가 보유하고 있는 SUN 2000E DataCenter 서버에 설치한 SYBASE 를 데이터베이스 엔진으로 하여 구동한다. 실험 데이터는 데이터 u-pload 프로그램에 의해 실험장비 제어용 컴퓨터에서 네트웍으로 데이터 서버에 전송된다. 데이터 upload 프로그램은 웹 브라우저에서 사용할 수 있다. 실험자는 Internet Explorer 나 Netscape와 같은 웹브라우저에서 데이터 입력 페이지를 열고 실험에 관계된 사항들을 입력하면, 입력된 정보들은 과학전산실의 데이터 서버로 전송되어 서버의 데이터베이스에 추가된다. 이 때에 실험자의 컴퓨터에 저장되었던 실험데이터도 다른 정보들과 함께 데이터 서버에 전달된다. 이렇게 데이터가 저장된 순간부터, 분석 신청자는 데이터를 검색하여 분석프로그램을 실행할 수 있다.

데이터 분석 프로그램은 자바 (Java) 언어를 사용한다. 자바 프로그램은 사용자의 컴퓨터에서 실행되며, 프로그램에서 사용하는 데이터 파일은 URL 주소를 이용하여 사용자에게 전송된다. 원격 데이터 분석 소프트웨어의 개발에 자바 언어를 사용함은 프로그램이 각각의 클라



〔그림 1〕 원격 데이터 분석 시스템 구성도

이언트에서 실행됨으로써 서버에 과부하가 걸리는 것을 방지하며, 사용자들의 그래픽 환경에 무관하게 프로그램을 실행할 수 있다는 장점을 가진다. 현재의 시스템은 JDK (Java Development Kit) 1.1.7 버전 (Weiner, 1998) 을 사용하여 개발하였다.

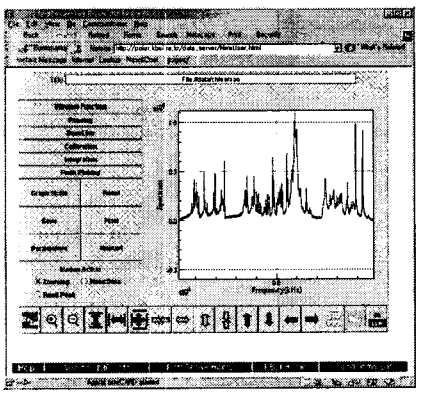
Ⅲ. 데이터 분석 시스템의 기능

데이터 분석 시스템의 프로그램들은 상용 데이터 분석 프로그램에서 많이 사용하는 주요 기능들을 그대로 수행하며, 이 때에 계산에 사용하는 실험 데이터는 사용자 컴퓨터의 하드디스크에 저장한 것이 아닌, 데이터 서버에 저장된 것을 네트웍을 통하여 전달받는다. 자바 프로그램은 사용자의 platform에 관계없이 프로그램을 실행할 수 있는데, 이는 연구자들간의 정보 공유에 있어서 커다란 장점으로 대두된다. 이러한 시스템의 특징을 고려할 때 자바 언어로 개발된 원격 데이터 분석 시스템은 통신 기술과 자바 프로그래밍 기술이 발전함에 따라 연구자들간의 정보 공유를 촉진하는 역할을 할 것이다.

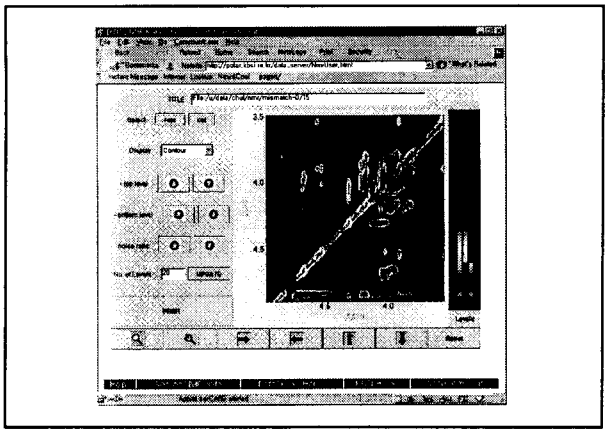
여기서는 5개 분석 장비의 데이터 분석 프로그램들에 대한 구체적인 설명과 함께 현재 개발된 기능들을 소개하기로 한다. 데이터 분석 프로그램은 기존의 상용 데이터 분석 시스템을 참조하여 제작하였으나 프로그램의 다운로드 및 수행 속도를 감안하여 연구자들이 많이 쓰는 필수적이고 중요한 기능만을 구현하도록 시스템 개발 방향을 설정하였다.

1. 핵자기공명 분광기 (NMR Spectrometer)

원자핵이 자기장 속에 놓이면 핵의 고유한 양자적 성질에 따라 일정 주파수의 전자기파를 흡수하거나 내어놓게 된다. NMR (Nuclear Magnetic Resonance) 분광기는 이와같은 공명 현상을 이용하여 분자의 구조나 운동성을 연구하는 장비이다. (Hoch, 1996) 현재 기초과학지원연구소의 대덕본소에서는 600MHz, 400MHz, 300MHz 의 NMR 분광기를 보유하고 있다. 자기공명 스펙트럼 데이터는 1차원 fid 데이터와 2차원 ser 데이터, 다차원 데이터로 크게 구분된다. NMR 데이터의 분석은 분석 방법에 따라 필요한 정보를 얻을 수도, 잃을 수도 있으므로 분석 목적에 따라 방법을 달리한다. 그러므로 분석지원데이터에 대한 원격 분석 시스템의 개발은 분석을 의뢰한 연구자가 네트웍을 이용하여 나름대로의 스펙트럼 분석을 수행할 수 있게 하여 분석지원의 질을 한 단계 높일 수 있었다. NMR 데이터 분석 프로그램은 현재



〔그림 2〕 1차원 NMR 데이터 분석 화면



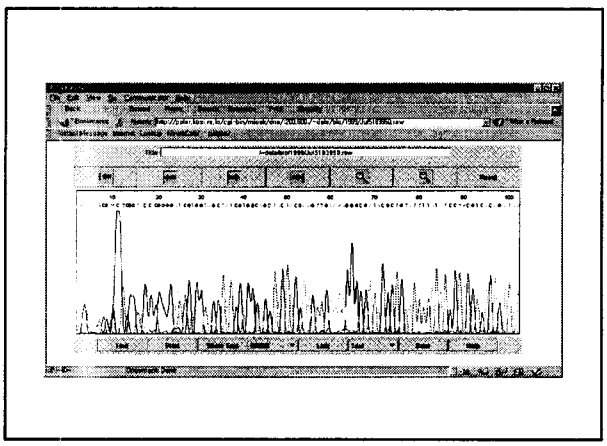
〔그림 3〕 2차원 NMR 데이터 분석 화면

실험장비에 연결되어 설치되어 있는 Bruker 사의 XWIN-NMR 소프트웨어의 기능 중 데이터 분석에 있어서 필수적이고 중요한 것들을 우선으로 개발하였다.

1차원 데이터의 분석은 스펙트럼의 변환에 중점을 두어, Apodization, Phase Correction, Baseline Correction, Chemical Shift 등의 다양한 메뉴를 제작하였으며, fid 데이터와 스펙트럼을 그래프로 볼 수 있다. 그래프는 마우스와 버튼을 이용하여 확대, 축소, 이동할 수 있으며, 스펙트럼의 Peak 위치를 적거나, 마우스로 클릭한 구간의 면적을 계산하는 기능, 스펙트럼을 프린터로 출력하는 기능 등이 있다.

2차원 ser 데이터에 대해서는 스펙트럼의 그래픽 표현에 중점을 두어 프로그램을 개발하였다. 스펙트럼을 등고선으로 표시하거나 색깔로 구분하여 볼 수 있다. 등고선 준위의 갯수는 사용자가 임의로 정할 수 있다. 스펙트럼 값이 양수인 것 또는 음수인 것만 선택하여 그릴 수 있으며, 0 근처의 noise 를 걸러내는 메뉴도 있다. 2차원에서도 마찬가지로 마우스와 버튼을 이용하여 그래프를 확대, 축소 또는 이동할 수 있고, 결과를 프린터로 인쇄할 수 있다. 2차원 데이터는 가로 또는 세로 방향으로 단면을 자르면 1차원 스펙트럼을 얻게 되는데, 사용자가 마우스를 클릭하여 자르는 위치를 선택하면 그 부분의 1차원 스펙트럼을 독립된 윈도우에 그래프로 그리고 이를 인쇄할 수 있도록 하였다.

2. DNA Sequencer



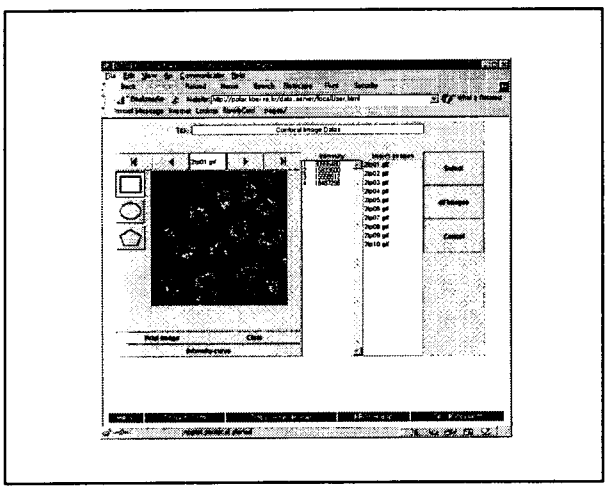
〔그림 4〕 DNA 염기서열 분석화면

DNA 염기서열(DNA Sequence) 그래픽 분석 프로그램은 기초과학지원연구소의 DNA 염기서열 분석 시스템(Sequencing Analysis System)인 377 DNA Sequencer에서 사용하고 있는 ABI PRISM DNA Sequencing Analysis Software의 기본 기능을 제공한다. 이 분석 프로그램은 염기별로 서로 다른 색의 스펙트럼을 기본 화면으로 와 같이 보여준다. DNA 염기서열을 스펙트럼을 화면 위 부분에 나타내고, 염기서열 중 바탕색을 다르게 하여 삽입(insert) 구간을 표시하였다. 버튼을 이용하여 스펙트럼

전 구간에 걸쳐 이동하며 분석할 수 있고, 마우스를 이용하여 관심 영역을 선택함으로써 스펙트럼을 축소 또는 확대할 수 있으며, 축소 및 확대 버튼을 이용하여 수평 방향으로의 축소 및 확대가 가능하다. 스펙트럼 전영역에 걸쳐 원하는 부분의 DNA 염기서열을 탐색할 수 있으며, 분석자가 원하는 형태(ABI 파일, 문서 파일, 스펙트럼 파일)로의 데이터 파일 저장이 가능하고, 인쇄 시 인쇄용지 크기에 맞게 스펙트럼 화면을 축소하여 인쇄한다. DNA 염기서열을 스펙트럼 형태와는 별도로 문서형태로 볼 수 있으며, 염기서열을 GenBank 등과 같은 대규모 GENOM DB 시스템과 연결하여 검색할 수 있는 기능 등이 있다.

3. 현미화상분석시스템 (Microscopic Imaging System)

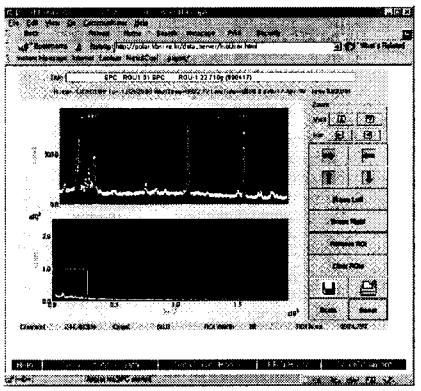
현미화상분석시스템은 레이저를 광원으로 사용하여 시료를 광학적 절편 형태로 관찰할 수 있는 형광현미경의 일종으로 살아있는 세포의 실시간 분석과 세포내 소기관 또는 신호전달물질의 시간적 변화를 관찰할 수 있다. 현미화상 분석시스템의 실험 데이터는 TIF, JPG 또는 GIF format 의 그림 파일들로 구성된다. 이 데이터들은 시료를 일정한 시간간격으로 촬영한 사진들로서 한 화면에서 각 부분의 광도를 비교하고, 일정 부분의 광도가 시간에 따라 어떻게 변화하는지를 계산해야 한다. 이렇게 계산된 광도의 변화량은 파일에 저장하거나 그래프로 표현하여 인쇄한다.



〔그림 5〕 현미화상분석시스템의 화상데이터 분석 화면

데이터 분석 프로그램은 크게 세 부분으로 나뉜다. 첫째는 데이터 파일을 선택하는 모듈로, 데이터베이스에 저장된 그림 파일들을 차례로 보고 이 중에 관심있는 파일들만을 선택하여 파일 목록을 재구성한다. 이렇게 파일 목록이 구성되면 화살표 버튼에 의해 그림 파일을 차례로 볼 수 있다. 둘째는 하나의 사진에서 여러 부분의 광도를 비교하는 모듈로, 사각형, 타원, 다각형 중의 한 모양으로 관심 영역(Region of Interest : ROI)을 선택하여 광도를 계산할 수 있다. ROI 는 원하는 모양의 버튼을 누른 뒤 해당 지점을 마우스로 당기어 선택하며, ROI 가 정해지면 해당 영역의 광도가 목록에 추가된다. 세째 모듈은 화면의 일정부분을 ROI 로 선택하여 그 부분의 시간에 대한 광도 변화를 계산하고 비교하는 부분이다. 광도의 변화는 별도의 윈도우를 생성하여 1차원 그래프로 표현되며, Print 버튼을 누르면 이 그래프를 인쇄할 수 있다. 일정 부분의 시간에 대한 광도 변화는 ROI 가 다각형인 경우에 계산 시간이 다소 오래 걸리는 단점이 있다. 이 모듈은 알고리듬의 개선에 의해 보완되어야 한다.

4. 자연방사능 측정기
(Natural Radioactivity Measurement System)



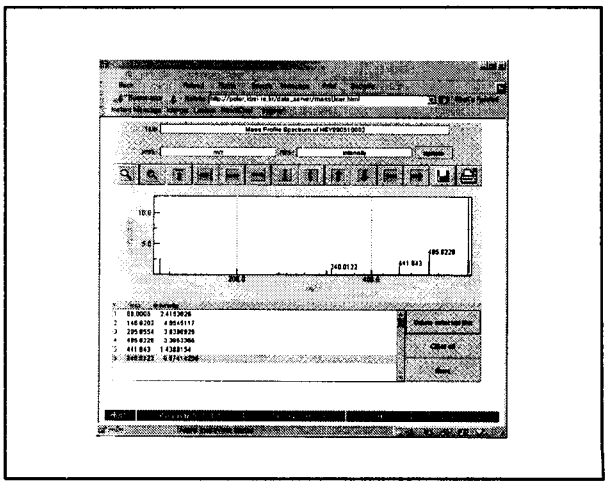
〔그림 6〕 자연방사능 측정기에서 얻어진 방사선 분포 데이터의 분석 화면

자연방사능 측정기는 방사선이 가지는 에너지를 전기로 변환하여 방출된 방사선의 에너지와 갯수를 측정함에 의해 방사선 방출 핵종을 알아내고, 방사선의 방출량을 측정한다. 자연방사능 측정 데이터는 MCA와 SPC 라는 두 가지 형식의 데이터로 저장된다. 각 데이터는 실험 데이터 저장 구조에는 차이가 없으나 실험 관련 파라미터들이 파일 앞부분에 서로 다른 구조로 저장되어 있다. 따라서 파일을 읽는 부분에서 파일의 종류에 따라 자바 프로그램을 달리 사용하였다. 또한 입력된 파라미터의 종류에도 차이가 있어서 분석 프로그램의 판넬에 약간의 차이가 있다. 그러나 기본 데이터 분석 방법은 동일하다.

자연방사능 측정 프로그램에서는 두 개의 그래프를 한 화면에 보여주었다. 아래 부분은 전체 스펙트럼을 나타내며, 그 중 직사각형으로 표시한 부분을 확대하여 윗 그림에 나타내었다. 마우스을 당겨서 아래 그래프의 직사각형을 이동하면 윗 부분의 스펙트럼이 이동한다. 마우스를 움직여서 직사각형의 크기를 확대 또는 축소하면 위 그래프의 스펙트럼은 이에 따라 축소 또는 확대된다. 스펙트럼의 확대, 축소, 이동은 오른쪽 판넬에 있는 버튼들을 이용하여서도 가능하다.

오른쪽 판넬의 그래프 이동 버튼 아래에는 ROI (Region of Interest) 관련 버튼들이 만들어져 있다. 자연 방사능 측정 데이터에서는 각각의 Peak 에 대한 적분값이 중요하여, Peak의 위치 및 세기에 따라 그 시료에 어떤 종류의 방사능이 측정되었는가를 판별할 수 있다. 마우스로 ROI 를 선택하거나 삭제, 수정함으로써 Peak 들에 대한 정보를 얻을 수 있다. 선택된 ROI는 화면에 빨간 색으로 표시된다. ROI 관련 버튼들 아래에는 파일의 저장, 인쇄, 초기화 버튼들이 준비되어 있다.

5. 고분별능 탄뎀 질량분석기
(High Resolution Tandem Mass Spectrometer)

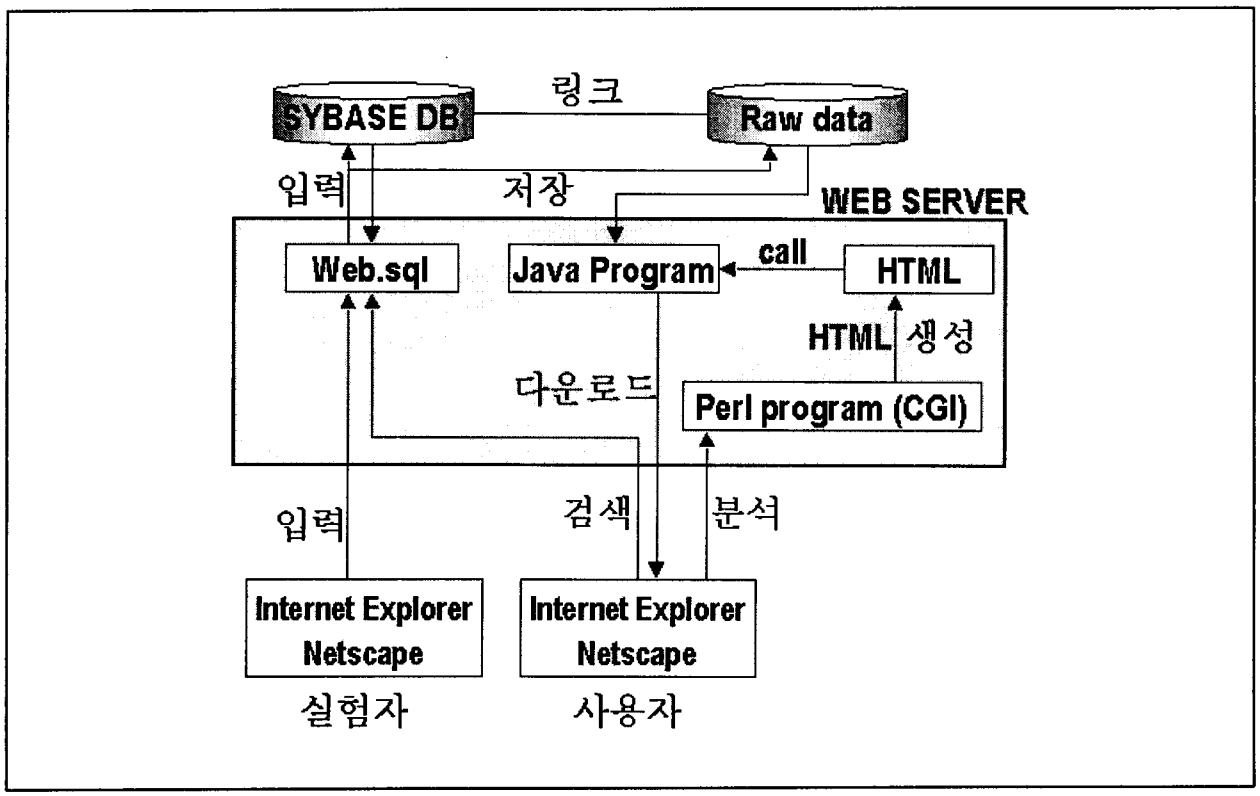


〔그림 7〕 탄뎀 질량분석기의 스펙트럼 분석 화면

탄뎀 질량분석기는 전기장과 자기장을 이용해 이온의 질량을 측정하는 질량분석기로서 High Energy Collision을 이용한 탄뎀기법으로 생체고분자 물질의 구조분석과 물질의 원소조성에 대한 정보를 제공한다. 질량분석시의 스펙트럼 분석 프로그램은 다른 장비들에 비해 비교적 단순한 구조를 가진다. 현재 개발된 프로그램에서는 스펙트럼 중에서 관심있는 Peak의 위치를 마우스로 클릭하면 해당 peak의 강도를 그래프에 적고 이를 peak 목록 란에 추가한다. Peak의 목록 중에 일부를 삭제하려면 목록의 해당 칸을 클릭한 후 Delete 버튼을 누른다. Save 버튼은 peak 목록을 파일로 저장하는 기능을 한다. 다른 분석 프로그램들과 마찬가지로 질량 분석 스펙트럼도 그래프의 확대, 축소, 이동, 저장, 인쇄 기능이 있다.

Ⅳ. 프로그램 내부 구조

웹 브라우저를 이용하여 네트웍으로 데이터를 입력, 검색하고 분석하는 원격 데이터 분석 시스템은 SYBASE DB 엔진과 SYBASE의 웹 인터페이스용 소프트웨어, 웹에서 실행하는 프로그램인 CGI (Common Gateway Interface) 프로그램 (Boutell, 1996), 자바 언어로 개발한 데이터 분석 프로그램이 복합적으로 사용된다. 프로그램의 내부 구조는[그림 8]과 같다. 실험자가 웹 브라우저에서 실험 데이터 관련 사항을 입력하면, 입력한 정보는 web.sql 이라는



〔그림 8〕 원격 데이터 분석 시스템의 프로그램 내부 구조

SYBASE의 웹 인터페이스 소프트웨어를 이용하여 데이터베이스에 입력된다. 이 때 실험 데이터 자체는 데이터베이스에 링크된 채로 일반 파일 시스템으로 저장된다. 외부 사용자가 실험 데이터를 검색하는 것도 역시 웹 브라우저를 이용한다. 웹 브라우저에서 키보드를 이용하여 실험에 대한 접수번호와 접수일을 입력하면 이 정보는 web.sql 을 거쳐서 SYBASE 에서 검색되고, 그 결과는 HTML 형식으로 출력되어 사용자에게 보여진다.

검색된 결과에서 자바 프로그램으로 데이터를 분석하려면 웹 화면에서 Analysis 를 클릭하는데, 이 때에 CGI 프로그램이 동작하여 자바 프로그램을 부르는 HTML 파일을 임시로 생성한다. 현재는 Internet Explorer 와 Netscape 에서의 자바 프로그램 호출 방법이 다르므로 두 가지 경우를 모두 적은 HTML 파일을 생성한다. 이 때에 SYBASE에서 검색된 실험 데이터 파일은 자바 프로그램의 실행을 위한 파라미터 값으로 설정된다. CGI 프로그램에 의해 생성된 HTML 파일에 의해 자바 프로그램은 사용자에게 다운로드되어 사용자의 PC에서 실행되고, 여기에 파라미터로 설정된 데이터 파일이 네트웍으로 전송되어 사용자는 원하는 데이터를 그래프로 볼 수 있게 된다. 자바 프로그램과 데이터 파일을 다운로드 받은 후의 모든 프로그램 실행은 사용자의 PC에서 이루어지므로 프로그램 실행에 의해 네트웍에 부하가 걸리지는 않는다.

이와 같이 원격 데이터 분석 시스템은 현재 많이 사용되고 있는 웹 기술들 중에 편리한 부분들을 경우에 따라 채택하여 통합적으로 운영함으로써 다양한 데이터 서비스를 구현하는 데에 중점을 두었다. 자바 언어는 현재 C 와 같은 일반 언어보다 실행 속도면에서는 뒤지지만 컴퓨터의 기종에 관계없이 그래픽 처리를 할 수 있다는 장점으로 애용되고 있으며, 자바 기술이 발전함에 따라 계산 속도도 급격히 향상하리라 기대한다. 원격 데이터 분석 시스템은 자바 서브렛 (Servlet) 이라는 개념을 도입하여 시스템의 성능과 기능을 발전시킬 계획이다.

V. 데이터 구조

원격 데이터 분석 시스템의 데이터는 각 장비에 따라 다른 format 으로 저장된다. 1차원 NMR 데이터는 fid format의 실험 데이터를 실수부와 허수부로 구분하여 HDF format (http://hdf.ncsa.uiuc.edu/)으로 저장하여 데이터의 관리를 쉽게 하였다. 각 데이터는 4-byte integer 타입으로 저장된다. 2차원 NMR 데이터는 스펙트럼 데이터인 rr, ri, ir, ii 데이터를 각각 따로 저장하였다. 2차원의 경우에는 대부분의 데이터 분석이 rr 파일에 대해 수행되므로 데이터 전송과 데이터를 읽는 데에 걸리는 시간을 줄이기 위하여 파일을 분리하였다.

탄뎀 질량분석기의 스펙트럼 데이터는 .jpf 로 명명되는 JCAMP-DX format의 ASCII 데이터 파일로 저장하였다. 원래의 raw data는 JEOL-DX format으로 실험 관련 파라미터들이 여러 파일들에 산재되어 있고 복잡한 구조를 가지고 있다. 질량분석기의 소프트웨어에서는 데이터를 JCAMP-DX format 과 HP-GL format

으로 변환하여 저장하는 기능이 있는데, JCAMP-DX format 은 실험 관련 파라미터들을 헤더 부분에 기록하고 뒷부분에 스펙트럼 데이터를 기록하는 간단한 구조를 가진다. HP-GL 은 그래픽 format 으로 스펙트럼을 인쇄할 때에 사용한다. 단순히 질량분석 스펙트럼만을 분석하는 데에는 JCAMP-DX format 의 .jpf 파일이 JEOL-DX 보다 더 효과적일 수 있으므로 데이터베이스에는 .jpf 파일들을 저장하였다.

DNA Sequencer의 데이터는 ABI PRISM 이라는 DNA 서열 분석 소프트웨어에서 사용하는 ABI format 의 데이터를 저장하였다. ABI format 역시 헤더 부분에 데이터의 offset 위치를 지정하고 파일의 뒷부분에 데이터를 binary 로 저장하는 구조를 가진다.

현미화상 분석시스템에서는 지금까지 주로 TIF format의 파일들을 사용하였으나, Java 1.1.7 버전에서는 GIF 또는 JPG format 만을 지원하므로 데이터를 GIF format 으로 변환하여 저장하였다. 한 실험에 대해 여러 개의 그림 파일들이 생성되므로 한 실험에 대한 파일들을 한 디렉토리에 모아서 저장하는 방법으로 파일 시스템을 구성하였다. Java 1.2 이상에서는 TIF format을 위한 함수들이 개발되어 있으므로, 이를 적용하면 TIF format의 파일들을 저장할 수 있게 된다.

자연 방사능 측정 데이터는 측정 장비에 따라 MCA와 SPC 라는 두 가지 데이터 format을 가진다. 두 format은 거의 비슷한 구조를 가지나 기록한 파라미터의 종류 및 데이터 위치에서 차이가 난다. 데이터베이스에는 각각의 format을 그대로 유지하여 파일을 저장하였으며, 데이터 분석 프로그램도 데이터 format에 따라 달리하였다.

지금까지 개발한 시스템들의 실험 데이터를 비교해 보면 각 장비들이 서로 다른 구조를 가지고 있으나 몇 가지 공통점을 발견할 수 있다. 우선 데이터의 전체 구조는 헤더와 데이터로 구분되며, 헤더 부분에는 실험에 대한 설명들이 일정 위치에 기록되고 데이터의 위치와 크기가 포함되어 있다. 데이터는 헤더에서 지정한 위치에서부터 기록되고 헤더로부터 데이터의 크기를 읽으면 데이터 값들을 읽을 수 있다. 이러한 공통 구조를 활용하면 여러 장비에 공동으로 사용하는 표준 데이터 format을 정의할 수 있다. 만일 표준 데이터 format 이 일반적으로 널리 알려져 있는 것이라면, 저장된 데이터를 상용 소프트웨어에서 읽고 분석할 수 있게 되어 데이터의 활용도를 높일 수 있다. 이에 표준 데이터 format 으로 미국 NCSA (National Center for Supercomputing Applications) 의 HDF (Hierarchical Data Format) 를 검토하고 있다. HDF format은 IDL, PV-WAVE 와 같은 상용 데이터 분석용 그래픽 소프트웨어에서 사용이 가능한 데이터 format 이다. 상용 데이터 분석 소프트웨어는 장비 특유의 데이터 분석을 위해서는 복잡한 프로그래밍이 필요하지만, 대용량의 데이터를 단시간 내에 처리할 수 있는 장점을 가지므로 연구 분야에 따라서는 이러한 소프트웨어의 사용이 필요하다.

한편 실험 데이터들이 DB에 축적됨에 따라 대용량의 데이터를 저장하기 위한 방안이 고려되어야 한다. 위에서 언급한 HDF format 과 같은 경우에는 데이터 압축 파일도 동시에 지원

하고 있음을 볼 수 있는데, 이와 같이 대용량 데이터에 대한 압축 처리 문제도 앞으로 연구되어야 할 것이다. 데이터의 압축은 사용자가 데이터를 다운로드 받는 데에 걸리는 시간을 줄일 수 있고 데이터 저장장치도 절약하는 효과가 있다. 단, 압축된 데이터는 압축을 푸는 데에 걸리는 시간을 감안하여 데이터 구조별로 최적의 방법을 선택하여야 한다. 데이터 파일의 압축 외에도 방대한 데이터의 처리를 위해 실험 데이터의 사용 빈도에 따른 저장 방법의 차별화를 고려하고 있는데, 예를 들어 6개월 이전의 데이터들은 DAT 테이프나 CD-ROM 에 저장하여 서버의 부하를 줄이는 방법이 있다. 오래 된 데이터는 사용자의 요청 시에만 서버에 mount 하여 검색할 수 있도록 함으로써 검색 빈도가 높은 데이터만을 서버가 지속적으로 관리하도록 한다.

Ⅵ. 결 론

지금까지 기초과학지원연구소의 과학전산실에서 개발한 원격 데이터 분석 시스템의 구조와 사용 방법 등을 기술하였다. 아직은 시스템의 성능 및 기능 면에서상용 데이터 분석 시스템에 미치지는 못하나, 네트웍을 통해서 데이터를 받아 처리할 수 있다는 점과 널리 사용되고 있는 웹 브라우저에서 데이터를 분석할 수 있다는 점은 장비의 사용자 계층을 확장시키는 데에 커다란 기여를 할 수 있다. 뿐만 아니라 실험 결과에 대한 가공을 사용자가 임의로 할 수 있게 되어, 결과 분석이 용이해졌다. 실험 데이터의 DB화도 본 시스템의 중요한 역할이 될 것이다. 고가의 장비로 수행된 실험들이 단 한 장의 그래프만을 남긴 채로 사장되지 않고 데이터베이스에 의해 정리 분류되어 관리됨으로써, 연구 결과들을 raw data 형태로 보존하게 되었다.

데이터베이스에 저장되는 데이터 format은 앞으로 상용 데이터 분석 tool 에서도 자유롭게 데이터를 읽고 분석할 수 있는 범용적인 데이터 format으로 변환할 계획이다. 기기별로 다양한 raw data의 format을 통일하여 연구소의 표준 데이터 format 을 설정하고, 연구자들이 이를 다운로드 받아 다양한 소프트웨어를 활용하여 분석할 수 있도록 제공할 계획이다. 현재는 미국 NCSA (National Center for Supercomputing Applications) 에서 개발한 HDF (Hierarchical Data Format) format 을 연구소의 표준 데이터 format으로 고려하고 있다.

원격 데이터 분석 시스템은 앞서 언급한 원격 공동연구 시스템의 기초 작업으로서 연구자들이 기초과학지원연구소의 데이터베이스를 공동으로 활용할 수 있는 바탕을 마련하였다. 여기에 원격 화상회의 시스템과 응용 프로그램 공유 시스템을 추가하여 원격 공동연구 시스템의 1단계 작업을 완료할 계획이다. 이렇게 구축된 시스템은 연구소의 네트웍 업그레이드 및 국가 초고속 통신망의 활용으로 on-line 분석지원, 공동연구를 활발히 추진할 수 있는 기반 시설의 역할을 할 수 있을 것이다.

본 연구는 기초과학지원연구소의 기관고유사업인 첨단기기 분석지원 사업의 세부과제로 진행되었다.

參 考 文 獻

권경훈, 최인식, 김미옥, 김준일, "기초과학지원 연구소 데이터 서버 구축", 제2회 과학기술 정보 워크샵, 1997.

Bair, R., "Collaboratories : Building Electronic Scientific Communities", *Chemical Sciences Roundtable Workshop: Impact of Advances in Computing and Communications Technologies on Chemical Science and Technology,* November 1998.

Boutell, T., *CGI Programming in C & Perl,* Addison-Wesley, 1996, pp.13-20.

Casper, T.A. and et al., "Remote Experimental Environment: Building a Collaboratory for Fusion Research", *Computers in Physics,* Vol.12, No. 3, May/Jun 1998.

Fredian, T.W. and Stillerman, J.A., "MDSplus Remote Collaboration Support - Internet and World Wide Web", *Fusion Engineering and Design,* Vol. 43, 1999.

Hoch, J.C. and Stern, A.S., NMR-Data *Processing*, Wiley-Liss, Inc., 1996, pp.34-76.

Mercurio, P.J. and et al., "The Distributed Laboratory: An Interactive Visualization Environment for Electron Microscopy and 3D Imaging", *Communications of ACM,* Vol.35, No.6, June 1992.

Parvin, B., "Distributed Virtual Microscopy and Microcharacterization", *LBNL Pub.* #041415, May 1997.

Weiner, R.S. and Asbury, S., *Programming with JFC,* Wiley Computer Publishing, 1998, pp.1-5.