MGrid: 분자 시뮬레이션 그리드 시스템

(MGrid: A Molecular Simulation Grid system)

정 갑 주 [†] 이 종 현 [††] 조 금 원 [†††] 정 선 호 [††††]
(Karpjoo Jeong) (Jong Hyun Lee) (Kum Won Cho) (Seunho Jung)
황 선 태 [†††††] 허 대 영 [††††††] 최 영 진 [†††††††]
(Suntae Hwang) (Daeyoung Heo) (Youngjin Choi)

요 약 본 논문에서는, MGrid 시스템과 이를 통한 응용 어플리케이션으로서 Glycoconjugates 시뮬레이션 데이타베이스 구축에 대한 연구를 소개한다. MGrid 시스템은 분자 시뮬레이션 계산 및 분석에 대한 그리드 서비스를 상호 운용 가능한 방법으로 제공하는 그리드 시스템이다. e-Glycoconjugates는 당접합체류의 분자 시뮬레이션을 수행하는 그리드 포털이다. 이 프로젝트는 MGrid 시스템을 통해 PDB와 같은 단백질 구조 데이타베이스 상에서 지금까지 알려진 2000 여개의 glycan chain들과 100 여개의 당접합체류에 대한 분자 시뮬레이션 데이타베이스 구축을 목표로 하고 있다. 본 논문에서는, MGrid 시스템의 목표와 시스템 아키텍처, 현재 시스템의 구현과 e-Glycoconjugates의 초기 결과를 기술하고자 한다.

키워드 : 그리드 컴퓨팅, PSE, 분자시뮬레이션

Abstract In this paper, we present the MGrid system and its application for the construction of the Glycoconjugates simulation database called e-Glycoconjugates. The MGrid system is an integrated molecular simulation grid system for computing, databases, and analyses. For e-Glycoconjugates, we have been constructing the simulation database for 2,000 glycan chains and 100 glycoproteins until 2008. In this paper, we present the goal, architecture, and current implementation status of the MGrid system, and e-Glycoconjugates.

Key words : Grid computing, PSE, Molecular Simulation, Glycoconjugates

1. 서 론

MGrid 프로젝트의 주요 목표는 다음과 같다.

- 계산 그리드 상에서 가능한 그리드에 효율적인 방법으로 대규모의 분산 High Performance 시뮬레이션 서비스를 개발한다. 이러한 계산 그리드는 지리적으로 분산된 전 세계의 여러 기관에 이질적인 컴퓨팅 자원으로 구성되어 있다.
- 가능한 그리드에 대한 지식 없이 쉬운 방법으로 응용 연구자들에게 대규모의 분자 시뮬레이션 기반 바이오 연구 수행을 지원한다. MGrid 시스템의 사용은 응용 연구자에게 기존 연구 방법론보다 연구 생산성을 획기적으로 높일 수 있다.

MGrid 프로젝트는 대용량의 컴퓨팅 자원의 공유, 가상 실험 능력 및 과학자들 간의 국제적인 협업을 가능하게 한다. 이를 통해 개발된 MGrid 시스템은 현재 KISTI 그리드 인프라상에서 동작 되고 있다. 이러한 그리드 인프라는 각 기관간 슈퍼 컴퓨터들과 HPC클러스터들로 구성되어 있다. 또한, 이러한 인프라상에서 구축된 MGrid 분자 시뮬레이션 서비스는 응용연구인 e-Glycoconjugates라 불리는 당접합체의 시뮬레이션 데이타베이스 구축을 지원하고 있다. 이를 통해 MGrid 시스템을 검증하고, 연구자들이 쉽게 당접합체에 대한 시뮬레이션 연구를 쉽고, 효율적으로 할 수 있도록 지원한다.

† 종신회원 : 건국대학교 인터넷&멀티미디어 공학부 교수
jeongk@konkuk.ac.kr
†† 학생회원 : 건국대학교 컴퓨터공학과
lejohy@gcslab.konkuk.ac.kr
††† 정 회 원 : KISTI 슈퍼컴퓨팅센터 슈퍼컴퓨팅응용지원팀장
ckw@kisti.re.kr
†††† 정 회 원 : 건국대학교 미생물공학과 교수
shjung@konkuk.ac.kr
††††† 종신회원 : 국민대학교 컴퓨터학부 교수
sthwang@kookmin.ac.kr
†††††† 학생회원 : 국민대학교 전산과학
dyheo@cs.kookmin.ac.kr
††††††† 정 회 원 : 건국대학교 생명분자정보학센터
ototo@konkuk.ac.kr
논문접수 : 2005년 9월 3일
심사완료 : 2006년 4월 13일

본 논문의 제2장에서는, 그리드 환경에서의 분자 시뮬레이션을 위한 사용자 요구사항에 대해 살펴보고, 3장에서는 요구사항을 적용한 MGrid 시스템 설계와 그에 대한 구성요소에 대해서, 4장에서는 MGrid 시스템 구현에 대해서, 5장에서는 MGrid 시스템을 통한 e-Glycoconjugates 초기결과와 구축현황을 기술한다. 마지막으로, 6장에서는 본 논문의 결론에 대해서 기술한다.

2. 그리드 환경에서의 분자 시뮬레이션 서비스 요구사항

분자 시뮬레이션 기반 실험을 효율적으로 지원하기 위해서는 다음의 몇 가지 요구사항이 필요하다.

- *긴 시간 수행되는 시뮬레이션 작업들에 대한 사용자 제어*

긴 시간 수행되는 작업들은 많은 컴퓨팅 자원과 시간을 필요로 하기 때문에, 과학자들은 시뮬레이션 진행 상황을 모니터링하고, 시뮬레이션 작업이 정상적인 상황이 아닐 때, 제어가 가능해야 한다.

- *그리드 환경에서의 레가시 시뮬레이션 소프트웨어의 효율적 지원*

과학자들은 신뢰성 있는 결과 데이타를 위해 증명된 몇 가지 소수의 시뮬레이션 소프트웨어 패키지를 사용한다. 따라서, 그리드 환경에서 이러한 시뮬레이션 소프트웨어들을 통합하고 효율적으로 지원 가능해야 한다.

- *과학자들로부터 그리드 환경의 복잡성을 효율적으로 숨김*

대부분의 과학자들은 그리드 기술에 대해 제한된 지식을 가지고 있으므로, 너무 많은 혼란없이 그리드 기술을 활용할 수 있도록 복잡성을 감추어야 한다.

- *연구 방법론의 개발과 많은 작업에 대한 동시 실행에 대한 환경 제공*

과학자들은 순차적인 방법으로 실험을 진행한다. 그러나, 그리드 인프라내에서, 이용 가능한 컴퓨팅 자원에 따라 동시에 많은 작업을 수행할 수 있어야 한다. 또한, 그리드 커뮤니티는 과학자들에게 많은 실험들을 쉽고 간단히 관리할 수 있는 환경을 제공해야 한다.

그러나, 이러한 바이오 연구그룹의 요구사항들은 범용적인 작업 수행을 요구하는 그리드 그룹의 요구사항들과 서로 상충된다. 일반적으로, 전통적인 4계층 그리드 서비스 구조는 그림 1에서 나타난다. 이 구조에서, Client PSE와 Server들은 어플리케이션들에 대해 많은 기능을 지원할 수 있다. 그러나, 그리드 미들웨어와 지역자원 관리 소프트웨어와 같은 미들웨어 계층들은 매우 간단하고, 제한된 기능만을 제공하고 많은 제약이 그 인터페이스간에 존재한다. 기능에 대한 다양성을 제공하기 위해서는, 강결합 클라이언트/서버 아키텍처 구조가 필요하나, 4계층 그리드 서비스 구조에서는 위와 같은 이유로 구현하기 매우 어렵다.

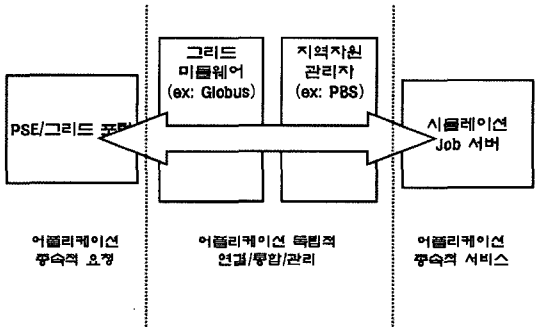

그림 1 그리드 기반 클라이언트/서버 시스템의 전통적 4계층 구조

전통적인 4계층 그리드 서비스 구조상에서 그리드 미들웨어인 Globus를 통해 작업을 제출하는 방법의 기술은 RSL(Resource Specification Language)을 통해 다음 표 1과 같이 이루어진다.

표 1 4계층 구조상에서의 작업 실행 RSL 예제

```
& (executable = a.out)
  (directory = /home/nobody )
  (arguments = input.txt output.txt)
  (count = 1)
```

이것은 PSE/그리드 포털이 원격지의 /home/nobody 상의 a.out을 실행시키되 인자로 input.txt와 output.txt를 사용하여 1번 실행시키는 일반적인 그리드 작업이다. 이러한 Globus Toolkit2를 통한 RSL 1은 범용적인 작업의 기술은 가능하지만, 분자 시뮬레이션 작업의 관리, 에너지 모니터링 등과 같은 어플리케이션 종속적인 기능의 수행은 매우 어렵다. 따라서 이러한 문제를 해결하기 위해서는 분자 시뮬레이션 관리를 위한 다른 구성요소가 필요하다.

3. 시스템 설계

3.1 설계 목표와 전략

사실, MGrid 시스템은 그동안 존재해왔던 분산 배치 시스템의 한 종류이다. 그러나 MGrid 시스템이 이러한 전통적 분산 배치 시스템들과 구별되는 설계상의 전략은 다음과 같다. 이것은 연구자들이 쉽고, 효율적인 분자 시뮬레이션 관리를 가능하게 한다.

- *싱글 시스템 뷰*

분산 컴퓨팅 기술에 대해 작은 지식을 가지고 있는 과학자들에게 여러 기관, 이질적이고, 분산된 플랫폼으

로 인한 시스템의 복잡성을 숨기는 것은 중요하다.

• *중앙 집중적 모니터링과 제어 제공*

과학자들은 계산 그리드상에 분산되어 있는 다수개의 동시 시뮬레이션 작업들에 대한 상태를 추적하고 제어하기를 원한다.

• *상호작용적 모니터링*

과학자들은 며칠에서 몇 달까지 걸리는 긴 수행시간이 걸리는 Simulation Job에 대한 모니터링을 하기를 원한다.

• *그리드 미들웨어상에 레가시 소프트웨어 통합*

CHARMM, AMBER, GAUSSIAN과 같은 레가시 분자 시뮬레이션 소프트웨어에 대한 자세한 기능을 다루는 것이 필요하다.

본 논문에서는, 2장에서 제시된 분자 시뮬레이션 서비스 요구사항을 충족하기 위하여 공유 인프라 기반 그리드 서비스 아키텍처를 설계하였다. 이것은 전통적인 그리드 기반 4계층 클라이언트/서버 아키텍처에 더하여, 시뮬레이션 관리를 위한 공유 인프라를 가진다. MGrid 아키텍처는 그림 2에 나타나고 있다. 공유 인프라 구조는 다음과 같이 설계되었다.

• *글로벌 ID 서비스*

모든 Job은 논리적인 ID를 가지고 임의의 시스템 컴포넌트상에서 글로벌하게 식별가능하다. 이러한 ID생성을 위해 각 기관간 시뮬레이션 Job 서버상에 분산 ID 관리자를 둔다. ID는 서버를 유일하게 식별하는 도메인 이름과 로컬 식별자(증가하는 수)를 사용하여 생성한다. (예:newcluster.konkuk.ac.kr-132)

• *분산 시뮬레이션 저장소*

Simulation Job이 제출될 때마다, 그것은 등록되고, 작업의 정보와 데이타는 각 시뮬레이션 서버의 공유 저장소안에 분산되어 저장된다. 이러한 저장소상의 Simulation Job에 대한 정보와 데이타는 어떠한 사이트에서든 ID를 통해 접근 가능하다.

• *그리드 포털*

이러한 인프라에 대한 서비스들은 웹브라우저를 통해 배치되는 그리드 포털을 통해 어디서든 이용가능하다.

이러한 인프라 상에서, 4계층 아키텍처(일반적인 그리드 서비스 아키텍처)는 그림 2의 상위에 위치된다. 본 논문에서 제시한 아키텍처와 4계층 아키텍처 사이의 다른 점은 본 논문에서 제시한 아키텍처는 오직 제어 정보의 통신(Globus GRAM이용)에 목적을 둔다는 점이다. 정보와 데이타의 접근(Globus GridFTP 이용)은 다른 시스템 컴포넌트에서 다루고 우리의 4계층 아키텍처에서는 데이타 접근을 다룰 필요가 없다. 이러한 아키텍처 설계는 시스템의 구현을 매우 간단하고, 쉽게 한다.

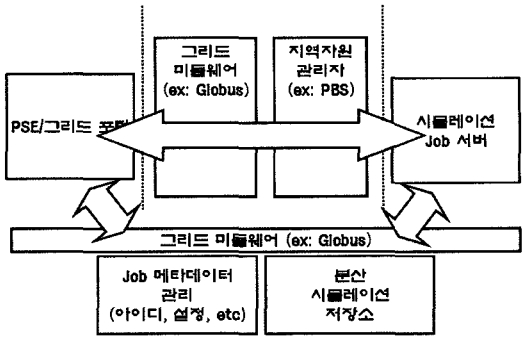

그림 2 공유 인프라 기반 그리드 서비스 아키텍처

공유 인프라 기반 그리드 서비스 아키텍처에서, 분자 시뮬레이션 작업은 3가지의 논리적인 단계(Job 등록, Task 추가, Task 시작)를 거쳐 수행된다. 이러한 제어에 관한 수행은 전통적인 그리드 서비스 아키텍처 상에서, 그리드 미들웨어(Globus의 GRAM)를 통해 이루어진다. 각 단계 제어 정보의 전달은 XML문서를 통해 이루어진다.

• *Job 등록(RegisterJob)*

PSE/그리드 포털이 Simulation Job 등록을 하게 되면 다음 표와 같이 분산 시뮬레이션 Job 서버내 작업 메타데이타 관리자는 시스템 식별자(ID)를 생성하고, 분산 시뮬레이션 저장소 내 ID와 같은 독립적인 작업 디렉토리를 생성한다.

표 2 Simulation Job 생성

```
def MakeJobObj(self, ReqMSG):
        JobUID = self.MakeJobUID()
        ReqMSG['JobID'] = JobUID
        WorkDir = self.WORKDIR + JobUID
        os.makedirs(WorkDir)
        ...
```

이것은 시스템내에서 유일하게 Simulation Job과 작업디렉토리에 매칭되며, 이러한 ID는 작업 메타데이타 관리와 분산 시뮬레이션 저장소를 통한 데이타의 접근에 사용된다. 시뮬레이션 서버간 Simulation Job Migration을 고려하여, 서버간 ID 매핑 테이블을 유지하고자 한다.

• *Task 추가(AddTask)*

PSE/그리드 포털은 등록된 Simulation Job에 응답받은 ID를 통해 Simulation Task를 추가한다. 작업 메타데이타 관리자는 Task를 만들고 응답하고, 이에 PSE/그리드 포털은 입력파일을 GridFTP로 작업 등록때 받은 작업 디렉토리상으로 파일을 전송한다.

• *Task 시작(StartTask)*

PSE/그리드 포털은 추가된 Simulation Task에 시작을 요청하면 분산 시뮬레이션 서버는 Task를 지역 자원관리자를 통해 수행한다. 그림 3은 이러한 시뮬레이션 작업수행 3단계인 Task 시작 XML 문서이다.

```
<?xml version="1.0" encoding="UTF-8"?>
<Message id="8f2dafa04e9b4d178077f996548a93d6"
         reply="https://newcluster.konkuk.ac.kr:8443/jss/projectmgr">
 <StartTask>
  <JobID>newcluster-771</JobID>
  <TaskID>newcluster-771-1</TaskID>
 </StartTask>
</Message>
```

그림 3 태스크 시작 XML문서

id 속성은 UUID로 구성된 메시지 식별자이며, reply 속성은 응답을 받기위한 Servlet주소를 나타낸다. <StartTask> 엘리먼트는 태스크 시작을 의미하며, <JobID> 엘리먼트는 Job을 구별하는 유일한 이름이자, MGrid 시스템 식별자(ID), <TaskID>은 태스크 식별자이다.

이러한 구조는 미들웨어 계층(그리드 미들웨어, 지역 자원관리자)을 통한 표준화된 작업 수행을 지원하면서도, 동시에 효율적으로 분자 시뮬레이션 서비스의 요구사항을 만족시킬 수 있다.

3.2 시스템 설계

MGrid 시스템은 (1)분산 시뮬레이션 서버, (2)MGrid PSE(Problem Solving Environment) / MGrid 포털 2개의 주요 컴포넌트로 구성된다.

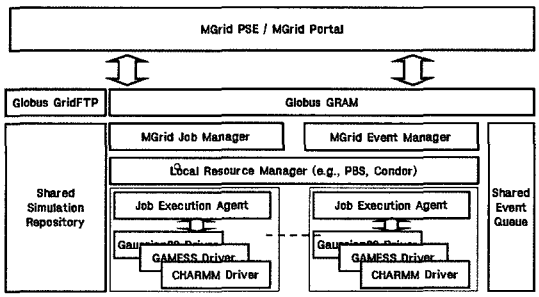

그림 4 MGrid 시뮬레이션 서버의 구조

3.3 분산 시뮬레이션 서버

MGrid 시뮬레이션 서버의 아키텍처는 그림 4에 나타난다. 이 아키텍처에서, 제어 정보에 대한 흐름은 MGrid PSE/MGrid 포털, 그리드 미들웨어(Globus GRAM), MGrid Job Manager/Event Manager, Local Resource Manager, Job Execution Agent, Legacy Driver로 이루어진다. 이러한 구조는 invocation과 제어 명령들에 대한 통신을 반영한다. 시뮬레이션 작업에 대해서, 중간 결과물이나 최종 데이타와 공유 이벤트 큐상의 이벤트에 대한 정보는 공유된 시뮬레이션 저장소에 의해 관리된다. 이러한 공유 시뮬레이션 저장소와 공유 이벤트 큐는 로컬 및 원격 사이트(PSE 혹은 그리드 포털) 둘 다 어떠한 시스템 컴포넌트에 의해서 접근 가능하다.

분산 시뮬레이션 서버의 디자인 초점은 다음과 같다.

• *객체 지향 설계*

각각의 컴포넌트는 객체로서 구현되고, 더 나아가 시뮬레이션 작업 정보와 데이타 또한 객체로서 구현된다. 시뮬레이션 작업정보는 Simulation Job, Simulation Task로 구성되어 있으며, 각각의 자료구조는 해당하는 메타데이타(작업이름, 작업설명, 소유자, 실행해야할 명령어, 실행해야할 소프트웨어 등)의 정보를 담고 있다.

• *범용 아키텍처 설계*

MGrid PSE 혹은 포털과 분산 시뮬레이션 서버 사이의 인터페이스들은 XML로 정의되므로, 다른 개발자들이 이러한 공개 XML 인터페이스를 정확히 지원한다면 이러한 컴포넌트의 교체 혹은 확장을 가능하게 한다.

• *레가시 소프트웨어에 대한 통합 지원*

JEA(Job Execution Agent)는 CHARMM, AMBER, GAUSSIAN, GAMESS와 같은 다양한 시뮬레이션 소프트웨어에 대한 통합 인터페이스를 제공하도록 설계된다. MGrid Job Manager와 JEA간 통합 인터페이스는 XML문서로 이루어진다. 표 2는 MGrid Job Manager가 JEA에게 Simulation Task를 시작하고자 명령하는 XML문서이다. 이렇게 구성된 XML정보는 JEA에 의해서 요청되는 소프트웨어에 따라 적절한 Driver를 호출하고, install-type과 version, command에 따라 각 기관에 설치된 시뮬레이션 소프트웨어 설치 경로에 동적으로 바인딩 한다.

표 3 MGrid Job Manager와 JEA간 Simulation Task 시작 XML 문서

```
<MSG>
<JOB-UID>newcluster-771</JOB-UID>
<TASK-UID>newcluster-771-1</TASK-UID>
<MSG-TYPE>STARTTASK</MSG-TYPE>
<TYPE>http://mgrid.or.kr/type/software/charmm/c28b2/l
 arge/charmm</TYPE>
<COMMAND>charmm &lt; mc.inp &gt; mc.out
</COMMAND>
<FILE></FILE>
</MSG>
```

• *시뮬레이션 결과에 대한 통합 접근 지원*

분산 시뮬레이션 저장소는 시뮬레이션 결과에 대해

유일한 URL기반 식별자를 할당하고, 이를 통해 GridFTP기반 글로벌 파일 접근을 제공한다. MGrid PSE와 포털은 Simulation Job을 생성할 때 응답받은 Job ID와 작업 디렉토리를 자료구조를 통해 GridFTP 프로토콜을 기반한 URL을 만든다. MGrid 시스템상에서 할당된 식별자는 PSE 혹은 그리드 포털 상에서 URL로 만들어지며, 공유 시뮬레이션 저장소상의 작업디렉토리로 매칭된다(그림 5).

• *지역자원 할당 정책 존중 설계*

MGrid Job Manager는 직접 시뮬레이션 작업을 지역자원에게 할당할 수 없다. 대신에 이것은 PBS, Condor와 같은 Local Resource Manager에게 자원 할당을 물어본다. 이러한 방법으로 MGrid 시스템은 각 기관의 로컬 자원 할당 정책을 존중한다. 이를 위하여 MGrid 시스템은 각 지역관리자에 대한 인터페이스를 제공한다. 표 3은 Local Resource Manager인 openPBS를 통해 JEA를 수행하는 스크립트이다.

MGrid Job Manager는 각 지역자원에 맞는 Local Resource Mananger(예, PBS)를 통한 스크립트를 자동으로 생성하고, 제출한다. 이후, Local Resource Manager는 자원 할당 정책에 따라 JEA를 기동한다. 이를 통해 기동된 JEA는 MGrid Job Manager에 등록한다. 이렇게 연결된 독립 채널을 통해 MGrid Job Manager는 JEA를 통해 수행된 시뮬레이션 작업에 대한 제어(상태 확인, 작업 시작/중지)와 상호작용적 모니터링(예. MD 모사를 위한 에너지, 크기, 각도계산 등)이 가능해진다. (그림 6)

표 4 OpenPBS를 통한 JEA 기동 스크립트

```
#PBS -l nodes=1
#!/bin/sh
/usr/bin/python2
/home/kgrid/kgrid001/new_LocalScheduler/WSM/WSM.
py newcluster-516
```

3.4 MGrid PSE와 MGrid 포털

MGrid PSE의 설계는 다음의 분자 시뮬레이션 모델에 기반을 두고 있다.

1) *Project.* 프로젝트는 명시적인 연구 문제에 대한 시뮬레이션 작업들의 모음이다.
2) *Simulation Job Array.* 시뮬레이션 작업 배열은 같은 스크립트 파일을 기반으로 하나, 다른 파라미터를 가진 시뮬레이션 작업들의 모음이다.

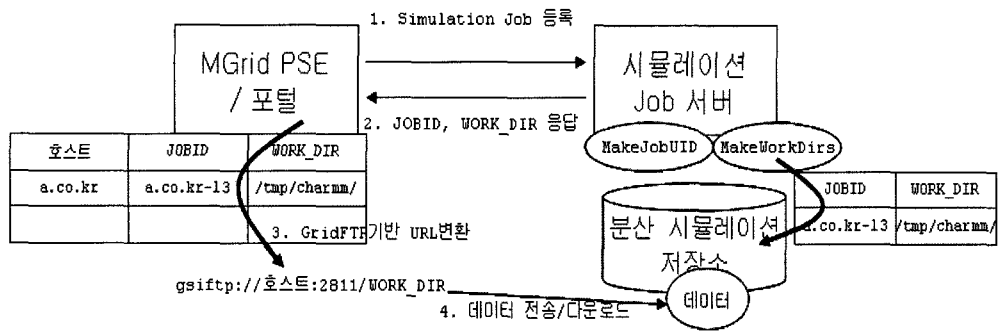

그림 5 Global 시뮬레이션 결과 접근

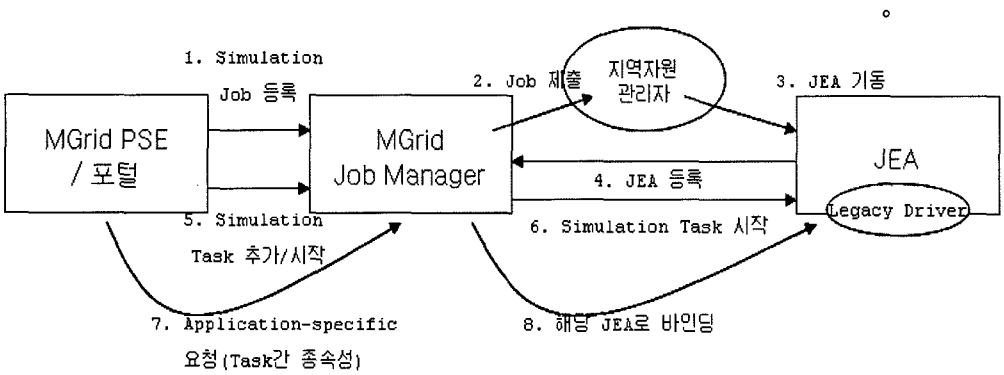

그림 6 상호작용적 모니터링

3) *Simulation Job.* 시뮬레이션 작업은 시뮬레이션 실험의 하나의 인스턴스로서 수행되도록 설계된 시뮬레이션 태스크들과 기타 태스크들의 그룹이다.
4) *Simulation Task.* 시뮬레이션 태스크는 시뮬레이션 코드의 하나의 실행이다. 예를 들어, CHARMM 시뮬레이션 스크립트의 실행을 나타낸다.
5) *Shell Task.* 쉘 태스크는 임의의 코드에 대한 하나의 실행이다. 이러한 태스크는 시뮬레이션 태스크로부터 파생된 출력 데이타의 포스트 프로세싱과 같은 다양한 기능을 수행한다.

가능한 실험 시나리오는 다음과 같다. 주어진 연구문제에 대해, 과학자는 1) 분자구조를 만들고, 시뮬레이션 스크립트를 작성한다. 이때 초기 파라미터 값도 결정된다. 이러한 활동으로부터의 결과들은 여러 Simulation Task들로 구성된 Simulation Job이다. 우리의 모델상에서, 시뮬레이션 태스크는 컴퓨터 시스템에 의해 직접적으로 수행될 수 있는 하나의 작업 단위이다. 2) 이 작업의 실행 후에, 과학자들은 다른 파라미터 값을 가진 다른 실험을 만든다. 그리고 나서, 그는 시뮬레이션 작업 배열을 만든다. 3) 이 작업 배열의 실행 후에, 많은 범위의 파라미터 값에 대한 시뮬레이션 결과를 모은다. 이러한 결과들은 실험의 일반적인 패턴 혹은 파라미터 값의 일반적인 경향을 찾는데 도움을 준다.

MGrid PSE/포털의 설계 초점은 다음과 같다.

• *추상화된 사용자 인터페이스 지원*

시뮬레이션 작업을 위한 자원의 현황이나 현재 상태, 소프트웨어 지원 등에 대해 알 필요 없이 과학자는 시뮬레이션 작업만 기술하고, 실행하면 된다.

• *어플리케이션-specific 지원*

MGrid PSE와 포털은 진행중인 시뮬레이션 작업들에 대한 상호작용 가능한 모니터링을 제공하고, 과학자들에게 관련된 결과 파일들에 대한 분석 혹은 시각화 도구를 통합하여 제공한다.

• *레가시 시뮬레이션 소프트웨어에 대한 통합 지원*

MGrid PSE와 포털은 과학자들에게 CHARMM, AMBER, GAUSSIAN과 같은 일반적인 분자 시뮬레이션 소프트웨어를 통합적인 인터페이스를 통해 다루는 것을 가능하게 한다.

MGrid 포털의 목적은 모든 MGrid 서비스들에 대해 centeralized한 접근성을 제공하는 것이다. 첫째로, MGrid 포털은 MGrid 공유 인프라에 대한 접근을 제공한다. 포털을 사용함에 의해서, 과학자는 시뮬레이션 작업에 대한 검색과 정보, 데이타를 구할 수 있다. 둘째로, 포털은 과학자들이 상호작용 가능한 방법으로 시뮬레이션 작업들에 대한 모니터링과 제어가 가능하다. 마지막으로, 포털은 MGrid 서비스와 인프라, 그리고 사용자에 대한 관리자 툴을 제공한다. 이러한 MGrid PSE와 포털의 주요기능은 다음과 같다.

• *분자 시뮬레이션 작업 워크플로우 지원*

분자 연구를 위한 시뮬레이션과 분석 작업은 필수적이다. 그러나 이러한 작업의 수행시간은 매우 분자의 크기, 연구목적에 따라 많은 시간과 복잡한 순서를 가진다. MGrid 포털은 이러한 Simulation Task간 의존성 정의와 실행 순서 정의를 통해 시뮬레이션 작업의 워크플로우를 지원한다. 그림 7은 MGrid Simulation Job을 이루는 Simulation Task간 워크플로우 정의를 나타내고 있다.

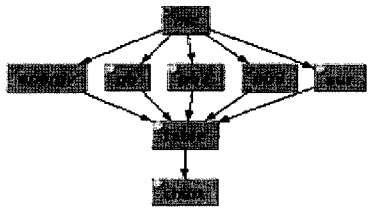

그림 7 MGrid 작업 워크플로우

Simulation Task는 사각형을 나타내며, 화살표는 각 Simulation Task간 실행순서를 나타낸다.

• *Simulation Job 복사 및 파라미터화*

분자 시뮬레이션 스크립트 내의 온도, 혹은 각도, 시뮬레이션 횟수 등 파라미터들은 시뮬레이션 결과를 결정하는 매우 중요한 역할을 한다. 연구자는 같은 작업에 대해 좀 더 좋은 연구결과를 위해 여러 다른 파라미터를 적용하는 것이 일반적이다. MGrid 시스템은 다시 작업을 정의하고, 파일을 업로드할 필요 없이 이미 포털상에 업로드된 작업을 복사하고, 새로운 파라미터를 변경하여 실험할 수 있도록 한다.

그림 8의 Parameters에서 해당 Task를 라디오 버튼으로 선택하고 Value를 변경한 후, Save버튼을 누르면 실험 파라미터 변경이 이루어진다.

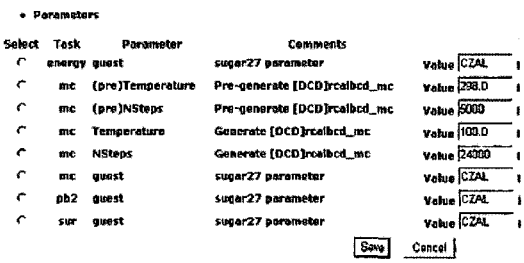

그림 8 작업 파라미터화

• *시뮬레이션 분석 도구 지원*

시뮬레이션 작업의 진행 중 작업의 정상적인 수행을 위한 에너지 그래프의 확인, 시뮬레이션된 작업 결과파일을 바탕으로 연구자에 최적화된 분석결과를 자동으로 엑셀파일로 변환해주거나, 연구자 컴퓨터상의 시각화 도구(예, gOpenMol)를 통합하여 분석에 사용한다. 그림 9는 시뮬레이션 작업에 대한 에너지 계산 그래프 화면을 나타낸다.

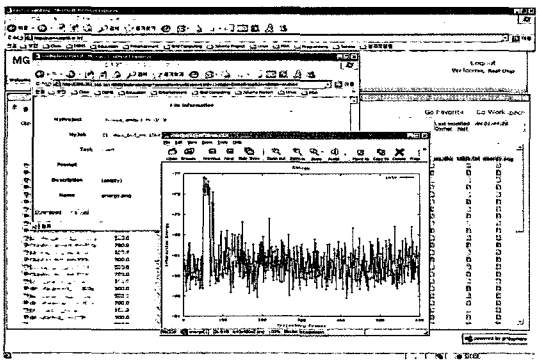

그림 9 MGrid 포털(시뮬레이션 분석)

4. MGrid 소프트웨어 구현

MGrid 소프트웨어 시스템의 프로토타입은 구현되어 현재 K*Grid 테스트베드상에서 수행중에 있다. MGrid 시스템을 구현하기 위해 사용된 주요 소프트웨어 시스템들은 다음과 같다.

• *Globus*. 주된 GRAM, GridFTP와 같은 Globus 미들웨어 컴포넌트는 본 논문에서 제시한 그리드 인프라를 만드는데 사용된다.

• *Gridsphere*. MGrid 포털을 만드는데 사용된다.

• *Python*. 이것은 분산 시뮬레이션 작업 서버를 개발하는데 사용된다. 본 논문에서는 빠른 프로토타이핑이 가능한 Python을 사용하였고, JAVA로 포팅을 계획하고 있다.

• *Eclipse*. MGrid PSE의 클라이언트 도구중 일부는 Eclipse에 의해 구현된 독립 어플리케이션이다.

5. 응용 프로젝트: e-Glycoconjugates

5.1 동기와 목표

2004년 단백질 치료제에 대한 시장은 약 1억 달러에 이르며, 그러한 단백질의 절반이상이 당단백질의 형태이다. 따라서, 당접합체의 구조 동역학에 대한 이해는 신약 개발의 전제조건이 된다. 따라서, 알려지지 않은 탄수화물 혹은 당쇄분자들의 형태를 예측할 수 있는 방법론 개발이 매우 중요하다. 이를 위해서는, 공용 구조 데이타베이스로부터 생물학적인 당쇄구조들을 모으고 분류해야 한다. 이러한 구조적인 모티프에 대하여 계산된 3차원 구조 라이브러리는 1차원 구조의 시퀀스 정보로부터 당접합체의 3차원 구조를 예측할 수 있는 강력한 도구를 제공할 수 있다.

하지만, 방사선 결정학이나 자기공명 분광기법과 같은 실험적인 방법을 사용하여 당접합체의 모든 구조정보를 얻기란 쉽지 않다. 그러나, 컴퓨터 시뮬레이션 기반 구조 분석은 이러한 실험적인 방법에 의한 제약을 받지 않는다.

e-Glycoconjugates의 연구 목표는 PDB를 포함한 구조 데이타베이스 상에서 나타난 2000여개의 당쇄분자들과 100여개의 당단백질들에 대한 생물구조학적 특성을 컴퓨터로 시뮬레이션하고, 그 결과를 데이타베이스화하는 것이다. 이러한 분자 시뮬레이션에 요구되는 CPU 시간은 하나의 Intel Xeon CPU를 기준으로 했을 때 약 60000일이 예상된다. 이것은 일반적인 컴퓨팅 자원으로 그러한 많은 양의 시뮬레이션 작업을 관리하는 것은 비현실적일 뿐만 아니라 합리적인 시간 내에 시뮬레이션 하는 것 또한 불가능하다. e-Glycoconjugates는 이러한, 하나의 컴퓨터로 다루기 불가능한 계산요구량을 충족시키기 위해 나타났다. MGrid의 고성능 컴퓨팅 인프라는 누구나 체계적인 당접합체 연구를 가능하게 함으로서 응용 연구의 큰 발전을 가져다 줄 수 있다.

5.2 그리드 기반 웹 포털

오늘날, 분자 시뮬레이션 연구자의 가장 큰 문제는 어떻게 사용자가 e-Glycoconjugates상에서 필요한 당접합체류에 대한 정보와 시뮬레이션 결과를 효율적으로 이용가능하게 만드는 것이다. 웹 포털은 이를 해결할 수 있는 가장 좋은 대안이다. 웹 포털은 e-business 포털을 만들고 배치시키기 위한 브라우저 기반 어플리케이션이다. 이것은 정보의 생성, 어플리케이션의 생성과 배치, 관리 환경을 포함하는 개방적이고 생산적인 도구 집합을 제공한다. 또한, 이러한 포털은 쉽게 커스터마이징이 가능하다. 따라서, 사용자는 자신의 연구 방향이나 영역에 따라 새로운 기능을 쉽게 추가가 가능하다.

e-Glycoconjugates 웹 포털은 이러한 웹 포털의 장점을 이용하여 데이타 수집 및 생성, 데이타의 제출 혹은 분석, e-Glycoconjugates 정보 데이타베이스와 시뮬레이션 데이타베이스 브라우징 구현에 사용하고, 사용자가 시뮬레이션 결과를 효율적으로 이용하고, 관리가 가능하도록 하고 있다. e-Glycoconjugates 웹 포털의 주요목적은 당접합체의 시뮬레이션 연구를 위한 계산 그리드와 데이타 그리드간 효율적인 공유 환경을 제공하는데 있다.

5.3 e-Glycoconjugates Simulation Job

e-Glycoconjugates 웹 포털상에서의 Simulation Job이란 동일한 분자구조를 가지는 당접합체의 3차원 구조 계산을 위해 수행하는 1개 이상의 Simulation Task로 구성되는 시뮬레이션 작업이다. 예를 들어, a-D-Gal-(1-3)-b-D-Glc-(1-4)-b-D-GlcNac 라는 당쇄분자의 3차원 구조를 계산하기 위해서는 최소 1개의 분자동역학 계산과 그에 대한 분석계산으로 에너지, 크기, 각도 계산등의 작업이 뒤따르며, 이들 각각의 Task가 하나의 Simulation Job을 이뤄 워크플로우로 수행될 수도 있고, 각 Simulation Task를 독립된 하나의 Simulation Job으로 만들어 수행 또한 가능하다. 이것은 연구자의 작업 편이성이나 계산의 효율성에 맞추어 사용자가 임의로 정할 수 있다.

5.4 e-Glycoconjugates 실험방법

당쇄분자는 단백질의 펩타이드 결합처럼 단일한 결합 방식으로 연결된 구조가 아니므로, 구성단량체가 같은 분자라 하더라도 화학적 결합방식의 종류에 따라 수많은 가지 수의 조합이 가능하다. 이렇게 다양한 경우의 수를 가진 당접합체의 구조를 연구하기 위해서는 가능한 많은 가상의 구조조합을 만들고, 이들을 하나의 Simulation Job 단위로 계산해서 공통적인 특징을 찾아야 한다.

e-Glycoconjugates는 MGrid상의 Job모델을 바탕으로, Project-Job Array-Job-Task 4개의 단위로 시뮬레이션 작업을 구조화하였다. 표 4는 현재 구축된 포털상에서 계산중인 당쇄분자를 Job 모델로 구조화한 사례이다.

4개의 단위 당으로 구성된 사당류(tetrasaccharide)가 하나의 Project가 되고, 사당류를 다시 분기점 구조가 있는지 없는지 여부에 따라 Job Array로 각각 구성할 수 있다. 이렇게 구성된 Job Array내에서 각 당쇄분자 화학 결합 방식에 따라 분자를 나누고, 이들을 각각 하나의 Simulation Job으로 만들어 시뮬레이션 작업을 수행한다. 이러한 Simulation Job의 반복적인 생성은 포털의 Job 복사의 기능을 통해 가능하다.

e-Glycoconjugates는 이러한 당쇄분자의 시뮬레이션 작업을 구조화된 틀 속에 넣어 분류하고, 각각의 태스크에 대하여 사용자가 부가설명을 입력하고 관리할 수 있는 웹 인터페이스를 제공하므로 사용자의 목적에 가장 적합한 시뮬레이션 접근법을 구현하는 데에 많은 도움을 준다. 또한 Job 복사와 파라미터화는 생물학적으로 유사한 구조를 가진 분자에 대한 반복작업에 대해 확장을 용이하게 해준다.

5.5 e-Glycoconjugates 구축현황

e-Glycoconjugates는 MGrid 인프라의 기술적인 지원으로 현재 각종 탄수화물 분자에 대한 3차원 구조 지도를 구축하고 있다. 표 5는 현재 시뮬레이션 시뮬레이션 작업이 수행 완료되거나, 수행 중 또는 계획중인 프로젝트에 대해 나타내고 있다.

현재까지 분석된 몇 가지 프로젝트의 의미는 다음과 같다. 1)Disaccharide를 다룬 Project에서는 당쇄분자의 결합방식과 그에 따른 구조 차이에 의하여 수용액 상에서 물분자와의 상호작용이 어떻게 달라지는가를 조사하였고, 특정 당쇄분자에서 강한 수소결합을 형성하는 것을 보임으로써, 당쇄분자의 생화학적 기능을 설명하였

표 5 당쇄분자에 대한 작업 구조화 사례

Project	Job Array	Job	Task
Tetrasaccharide	Linear	a-D-Manp-(1-3)-b-D-Manp-(1-4) -b-D-GlcNac-(1-4)-b-D-GlcNac	1)MD input 2)에너지 계산 3)크기 계산 4)각도 계산 5)Hydration 계산 6) Hydrogen Bond 계산
		a-D-Manp-(1-6)-b-D-Manp-(1-4) -b-D-GlcNac-(1-4)-b-D-GlcNac	
		b-D-Manp-(1-3)-a-D-Manp-(1-4) -b-D-GlcNac-(1-4)-b-D-GlcNac	
		b-D-Manp-(1-6)-a-D-Manp-(1-4) -b-D-GlcNac-(1-4)-b-D-GlcNac	
		a-D-Manp-(1-2)-a-D-Manp-(1-6) -a-D-Manp-(1-6)-b-D-Manp	
	Branched	a-D-Manp-(1-3)+ [a-D-Manp-(1-6) -b-D-Manp-(1-4)-b-D-GlcNac]	
		b-D-Manp-(1-3)+ [b-D-Manp-(1-6) -a-D-Manp-(1-4)-b-D-GlcNac]	
		a-D-Manp-(1-3)+ [a-D-Manp-(1-6) -a-D-Manp-(1-6)-b-D-Manp]	
		a-D-Manp-(1-6)+[a-D-Manp-(1-3) -b-D-Manp-(1-4)-a-D-GlcNac]	
		b-D-Manp-(1-6)+[b-D-Manp-(1-3) -a-D-Manp-(1-6)-b-D-Manp]	

표 6 프로젝트 현황

Project	작업 목표	작업 내용	진행 정도
Disaccharides	당분자 결합 종류별 특성 파악	이당류들의 구조와 수소결합 계산	본 계산 80종과 분석, 100% 완료
RNase	당접합체의 안정성 계산	Ribonuclease 단백질의 에너지 계산	본 계산 6종과 분석, 100% 완료
AFGP	당접합체의 부동성질 계산	결빙방지 당단백질의 구조 계산	본계산 10종, 50 % 완료
Tetrasaccharide	분기 구조의 영향 조사	사당류들에 대한 구조차이 계산	본 계산 40종, 30% 완료
DHFR	당접합체의 안정성 계산	Dihydrofolate 환원효소의 에너지 계산	본계산 5종 진행중, 20% 완료
Turn	당쇄결합의 구조적 영향 조사	Octapeptide에 대한 폴딩 구조 계산	본계산 4종 진행중, 10% 완료
Trehalose	당쇄분자 용액의 효과 계산	Trehalose 용액 내에서의 구조 계산	본계산 3종 준비중
Polysaccharide	거대당쇄분자의 구조 파악	각종 고분자당에 대한 수용액 구조 계산	본계산 100종 준비중

다. 2)RNase Project에서는 대표적인 단백질인 ribonucleaseA 효소에 글루코스아민 분자가 하나 치환 되었을 경우의 열 안정성이 증가하는 현상을 자유 에너지 계산과 비공유결합 능력 비교를 통하여 설명할 수 있었다. 3)Tetrasaccharide Project에서는 분기 구조를 가진 것과 가지지 않은 다양한 사당류에 대한 구조 계산을 통하여 당쇄분자 결합 종류와 분기구조가 전체적인 당쇄분자의 3차원 구조 형성에 어떠한 영향을 미치는지 조사중이다.

e-Glycoconjugates는 현재 수천 개의 당쇄결합에 대한 계산을 실행, 준비중에 있다. 이 시스템적인 접근방식은 대규모의 컴퓨팅 인프라만 지원된다면 일반적인 분자 시뮬레이션 방법으로 아직 알 수 없는 당접합체들의 전체 구조를 우리에게 알려줄 수 있을 것이다. 이를 통하여 그동안 실험적인 방법으로 구조가 알려지지 않았던 당쇄분자에 대해서도 구조 예측 서비스를 할 수 있는 데이타베이스를 구축할 수 있을 것이다.

6. 결 론

본 논문에서는 MGrid 시스템과 이를 이용한 당접합체 시뮬레이션 포털에 대해 기술하였다. MGrid 시스템은 현재 5개월 동안 운영 중이며, 당접합체의 시뮬레이션 데이타를 데이타베이스화하고 있다. 분자 시뮬레이션 기반 연구는 과학자들이 시뮬레이션 작업을 준비하고 실행하고, 이전 작업의 결과를 확인하고, 같은 소프트웨어 플랫폼상의 작업에 대한 분석을 수행할 수 있는 통합된 시뮬레이션 환경이 필요하다. 이러한 환경 없이 과학자들은 연구에 많은 시간과 비용이 들 뿐 아니라 실험 진행시나 파일 전송등과 같은 오류가 발생할 확률이 많아진다. MGrid 시스템은 이러한 과학자에게 충분한 컴퓨팅 자원과 사용하기 쉬운 통합 환경을 제공하도록 설계되었다. e-Glycoconjugates는 당접합체에 대한 분자 시뮬레이션 그리드 포털로 2007년까지 2000 여개의 당쇄분자와 100 여개의 당접합체에 대한 분자 시뮬레이션 결과를 데이타베이스화하고 데이타 서비스하는 것을 계획하고 있다.

참 고 문 헌

[1] Condor Project: http://www.condor.org
[2] GridLab Project: http://www.gridlab.org
[3] Globus Project: http://www.globus.org
[4] K*Grid Project: http://www.gridcenter.or.kr
[5] K*Grid Testbed: http://testbed.gridcenter.or.kr
[6] MGrid Project: http://www.mgrid.or.kr
[8] Foster, I., Kesselman, C., Tuecke, S.: The Anatomy fo the Grid: Enabling Scalable Virtual Organizations. Supercomputer Applications 15(3), 2001.
[9] Doucet, J.P., Weber, J.: J. Computer-Aided Molecular Design. Academic press, London 1996.
[10] Kim, H., Jeong, K., Lee, S. and Jung, S.: Molecular dynamics simulation of cyclosophoroheptadecaose (Cys-A). Journal of Computer Aided Molecular Design. Vol. 16, Issue 8-9, 2002.
[11] Varki, A.; Cummings, R.; Esko, J.; Freeze, H.; Harth, G.; Marth, J. Essentials of Glycobiology, *Cold Spring Harbor Laboratory Press* (1999).
[12] Roseman, S. Reflections on glycobiology, *J. Biol. Chem.* **276**(45), 41527-41542 (2001).
[13] Bertozzi, C.R.; Kiessling, L.L. Chemical glycobiology, *Science* **23**(291), 2357-2364 (2001).
[14] Woods, R.J. Computational carbohydrate chemistry: What theoretical methods can tell us, *Glycoconju. J.* **15**, 209-216 (1998).
[15] Jeong, K.; Kim, D.; Kim, M.H.; Hwang, S.; Jung, S.; Lim, Y.; Lee, S. A workflow management system and grid computing approach to molecular simulation-based bio-nano experiments, *Lect. Note. Compu. Sci.* **2660**, (2003).
[16] Choi, Y. Kim, D.W. Park, H. Hwang, S. Jeong, K. Jung, S. Prediction of chiral discrimination by b-cyclodextrins using grid-based Monte Carlo docking simulations. *Bull. Korean Chem. Soc.* **26**(5), 769-775 (2005).
[17] Weiner, S.J.; Kollman, P.A.; Nguyen, D.T.; Case, D.A. An all-atom force field for simulations of proteins and nucleic acids. *J. Comp. Chem.* **7**, 230-252 (1986).
[18] Van Gunsteren, W.F. Berendesen, H.J.C. GROningen MOlecular Simulation (GROMOS), library manual, *Groningen: Biomos BV.* (1987).

[19] Kale, L.; Skeel, R.; Bhandarkar, M.; Brunner, R.; Gursoy, A.; Krawetz, N.; Phillips, J.; Shinozaki, A.; Varadarajan, K.; Schulten, K. NAMD2: Greater scalability for parallel molecular dynamics. *J. Comp. Phys.* **151**, 283-312 (1999).
[20] Grant, G.H. and Richards, W.G. Computational chemistry, *Oxford University Press*, Oxford. (1995).
[21] Leach, A.R. Molecular modeling: Principles and Applications, *Academic Press*, London. (1996).

정 갑 주

1984년 2월 서울대학교, 컴퓨터공학과(학사). 1986년 2월 서울대학교, 컴퓨터공학과 인공지능(석사). 1996년 2월 New York University, Computer Science 박사 1995년 12월~1997년 8월 University of Florida, Post Doc. 1997년 8월~2001년 건국대학교, 컴퓨터 공학과 조교수. 2001년~현재 건국대학교, 인터넷&멀티미디어 공학부 부교수. 2006년~현재 BK21 u-Science 기반 신기술융합 사업단 단장. 관심분야는 Grid Computing, e-Science, Data Integration, 및 분산컴퓨팅

이 종 현

2003년 2월 건국대학교 컴퓨터 공학과 학사. 2005년 2월 건국대학교 컴퓨터 공학과 석사. 2005년 3월~현재 건국대학교 컴퓨터 공학과 박사과정. 관심분야는 Grid Computing, System Architecture, 분산컴퓨팅

조 금 원

1993년 인하대학교 항공우주공학과(학사). 1995년 한국과학기술원 항공우주공학과(석사). 2000년 한국과학기술원 항공우주공학과(박사). ~현재 KISTI 슈퍼컴퓨팅센터 슈퍼컴퓨팅응용지원팀장, KISTI e-Science 사업단 응용연구팀장(겸임) 관심분야는 슈퍼컴퓨팅, e-Science, 그리드

정 선 호

1985년 2월 서울대학교, 화학과(학사) 1987년 2월 서울대학교, 화학과(석사) 1993년 8월 Michigan State University, Biochemistry(박사). 1993년 9월~1995년 2월 Michigan State University, Post Doc. 1995년 3월~2003년 3월 건국대학교, 미생물공학과 조교수. 2005년 3월~현재 건국대학교, 미생물공학과 교수. 관심분야는 Bioinformatics, e-Science, Molecular modeling

황 선 태

1985년 서울대학교 컴퓨터공학과(학사) 1987년 서울대학교 컴퓨터공학과(석사) 1996년 Manchester University (PhD) 1997년~현재 국민대학교 컴퓨터학부 부교수. 관심분야는 e-Science, 그리드시스템, PSE, 공개소프트웨어

허 대 영

2004년 국민대학교 컴퓨터학부(학사) 2005년 국민대학교 전산과학(석사). 2006년~현재 국민대학교 전산과학 박사과정 관심분야는 그리드 시스템, 시스템 아키텍처, 디자인 패턴, 공개소프트웨어

최 영 진

1999년 2월 건국대학교, 미생물공학과(학사). 2001년 2월 건국대학교, 미생물공학과(석사). 2005년 8월 건국대학교, 미생물공학과(박사). 2005년 9월~현재 건국대학교, 생명분자정보학센터, Post Doc. 관심분야는 Biomolecular Simulations, High-performance computing, e-Science